Sciences naturelles avares en mots et sciences humaines en étalant trop ? Réponses statistiques à de vieux stéréotypes sur le discours scientifique

Marc Chalier[1,*], *Bettina* Eiber[1], et *Ursula* Reutner[1]

[1]Universität Passau, Innstraße 25, 94032 Passau, Allemagne

Résumé. Cet article présente une étude portant sur les différences de longueurs de phrases dans des articles de quatre disciplines scientifiques : la biochimie, la phonétique, la sociolinguistique et les études littéraires. La longueur de 10 phrases a été mesurée dans respectivement 20 articles de ces quatre disciplines, pour un total de 800 phrases. En plus de confirmer le stéréotype des phrases globalement plus longues dans les sciences humaines et plus courtes dans les sciences naturelles, il montre également des nuances à l'exemple de la linguistique. La phonétique, dont les méthodes sont souvent attribuables aux sciences naturelles, présente des phrases de longueurs statistiquement similaires à celles de la biochimie, alors que la sociolinguistique se rapproche pour sa part plus fortement des études littéraires. Nos résultats révèlent que la méthode de travail peut également influer sur la longueur des phrases : des faits observés empiriquement seront souvent présentés sur la base de phrases plutôt courtes, alors que des processus de réflexion favoriseront l'utilisation de phrases plutôt longues. Ces résultats montrent ainsi qu'une simple différenciation entre les disciplines n'est pas suffisante pour expliquer les différences observées, des nuances internes aux disciplines, explicables par des différences d'approches méthodologiques, devant également être prises en considération.

Abstract. **Natural sciences' word meagreness and 'humanities-babble'? Statistical answers to long-standing stereotypes about the scientific discourse**. This article presents a study on the differences in sentence length in articles from four scientific disciplines: biochemistry, phonetics, sociolinguistics and literary studies. The length of 10 sentences was measured in 20 articles per discipline, for a total of 800 sentences. Besides confirming the stereotype of overall longer sentences in the humanities and shorter sentences in the natural sciences, it also shows nuances based on the example of linguistics. Phonetics, whose methods are often considered to belong to the natural sciences, reveals statistically equally long sentences as biochemistry, whereas the sentence lengths found in the articles on sociolinguistics are closer to those of literary studies. Furthermore, our results show that the working method can also influence the sentence length: empirically observed facts are often presented on the basis of rather short sentences, while processes of reflection are more likely to result in longer sentences. These results show that it is not enough to differentiate between disciplines to explain the observed differences. Nuances within the disciplines, which can be explained by differences in methodological approaches, must also be taken into account.

1. Introduction

Au plus tard depuis Snow (1959), l'on a souvent opposé deux 'cultures scientifiques' dans le domaine du langage scientifique : celle des sciences naturelles, souvent représentées par la physique, et celle des sciences humaines, régulièrement représentées par le domaine des études littéraires (*cf.* Snow 1959 : 4). D'autres études suggèrent des cultures scientifiques conditionnées par différents contextes culturels, Galtung (1983 : 308) discernant les cultures scientifiques anglo-saxonne, germanique, romane et japonaise, alors que Clyne (1991 : 65) se concentre sur l'opposition entre un style qu'il attribue aux anglophones et un autre qu'il attribue aux germanophones. De telles démarcations, d'une part, entre des disciplines plutôt empiriques et des disciplines plutôt herméneutiques et, d'autre part, entre des cultures scientifiques liées à des espaces culturels, formulées d'autant plus de manière relativement unilatérale, relèvent en partie de vieux stéréotypes, qui se doivent d'être remis en question et qui devraient être mis à l'épreuve de données empiriques. C'est ce que Reutner (2008 ; 2009) s'est proposée de faire dans le cadre de deux études portant sur les représentations de l'idéal du langage scientifique de linguistes francophones (*cf.* Reutner 2008) et italophones (*cf.* Reutner 2009). Les résultats de ces questionnaires ont pu montrer que l'idéal de la simplicité fait quasiment l'unanimité parmi les informateurs·trices : 86,29% (107/124) des francophones et 83,61% (102/122) des italophones rejettent en

* Corresponding author : Marc.Chalier@uni-passau.de

effet l'emploi de phrases très complexes. D'importantes différences entre les langues et les cultures peuvent cependant être observées par rapport au style nominal, auquel l'on attribue la faculté de condenser les syntagmes et de supprimer l'agent. En effet, alors que ce style est favorisé par 56% des informateurs·trices francophones, il ne l'est que par 31% des informateurs·trices italophones. Un plus grand consensus est observable dans le cas de l'utilisation de participiales, qui représentent un autre moyen de condenser un texte. Les participiales sont en effet favorisées par 57% des francophones et 68% des italophones (*cf*. Reutner 2008 : 260*sq*. ; 2009 : 1412*sq*.).

Ces différents résultats et observations engendrent plusieurs questions que nous aborderons dans le présent article. L'on peut tout d'abord se demander dans quelle mesure la longueur des phrases est sujette à une variation intra- et interdisciplinaire : existe-t-il donc des différences entre les domaines de spécialité des auteur·e·s, et, si oui, pourraient-elles être expliquées par l'affiliation de ces auteur·e·s aux sciences naturelles ou aux sciences humaines ? Peut-on par ailleurs observer des différences au sein même des domaines de spécialité ? Dans le cadre de la linguistique par exemple, la phonétique est associée depuis Trubetzkoy ([1939] [7]1989 : 7) à des méthodes similaires non pas à celles des sciences humaines mais à celles des sciences naturelles. La question d'éventuelles différences dans le style d'écriture en phonétique et dans d'autres sous-disciplines linguistiques pourrait donc également présenter de l'intérêt. Celle-ci peut être étudiée à l'exemple de la longueur des phrases, en partant de l'hypothèse que cette longueur des phrases en phonétique pourrait effectivement avoir tendance à se rapprocher de celle qui est observable dans les sciences naturelles, alors qu'en sociolinguistique, par exemple, elle pourrait être plus proche de celle qui est observable dans les sciences humaines. Finalement, il serait intéressant de découvrir si la forte variation présente à l'intérieur de chacun des articles, qui avait été révélée dans les études précédentes (*cf*. p. ex. Rinck 2006), se confirme dans nos données.

Pour répondre à ces questions, la présente étude se propose d'étudier la longueur des phrases d'un point de vue quantitatif au sein d'un corpus de 80 articles originaux provenant de quatre disciplines : la biochimie médicale et clinique, la phonétique, la sociolinguistique et les études littéraires. Pour ce faire, nous dressons tout d'abord un bref état de l'art des aspects statistiques et variationnels de la longueur des phrases. Par la suite, nous présentons notre corpus et notre méthode d'analyse. Pour finir, nous décrivons les résultats quantitatifs obtenus et formulons quelques hypothèses pouvant expliquer les effets relevés.

2. État de l'art

Caractéristiques statistiques – La longueur de phrase est un trait linguistique dont les caractéristiques statistiques ont été décrites par la linguistique quantitative. Cette dernière définit la notion de « phrase » selon des critères graphiques. Selon Best (2005 : 300), par exemple, une phrase est une suite de mots graphiques introduite par une lettre majuscule et se terminant par l'un des signes de ponctuation forte (<.>, < !> et < ?>).[1] Elle est généralement relevée sur la base du nombre de syllabes, de morphèmes ou de mots par phrase (*cf*. Best 2005 : 300). Les longueurs des différentes phrases d'un texte peuvent varier d'un à plus de cent mots, la distribution des longueurs connaissant une très grande variation (*cf*. Barr 2001 : 377). La pertinence de mesures statistiques comme la moyenne ou la médiane est donc passablement limitée. La distribution montre généralement un taux relativement bas des phrases très courtes (de 1 à 5 mots) et très longues (de plus de 50 mots). La limitation de phrases courtes est explicable par la restriction des combinaisons possibles lors de l'utilisation d'un faible nombre de mots. La limitation des phrases très longues est, pour sa part, due au critère de la lisibilité (*cf*. Sigurd *et al.* 2004 : 50). Néanmoins, des réflexions ou des introspections peuvent se traduire par de longues phrases, qui expriment un flux de pensées ininterrompu (*cf*. Barr 2001 : 378). Mentionnons par ailleurs que l'alternance des longueurs de phrase crée un effet de contraste (*cf*. Barr 2001 : 377). Malgré ce problème de variation, certaines études ont cherché à déterminer des valeurs moyennes pour les longueurs de phrase dans certaines langues. Dans le cas de l'anglais écrit, Fengxiang (2007 : 129) indique une moyenne de 19,68 mots par phrase dans une partie du *British National Corpus* et 19,44 dans l'autre. Dans le cas du français québécois, Rouleau (2006 : 145) indique 23,8 mots par phrase dans son corpus de « textes généraux » provenant de quotidiens et de magazines québécois. Hoffmann (1998 : 417) relève une moyenne de 23,9 mots par phrase pour l'allemand.

Variation entre les traditions discursives – La longueur des phrases connaît une variation considérable dans différentes traditions discursives. Torttila/Hakkarainen (1990 : 34) relèvent une longueur de 7 à 9 mots par phrase dans les recettes de cuisine allemandes, un résultat s'avérant proche de ceux de l'oral et des annonces publicitaires. Ils expliquent ce résultat en invoquant la structuration du texte en des étapes de travail, qui sont exprimées à l'aide de phrases plutôt courtes. Les écrits scientifiques sont, au contraire, marqués par des phrases plus longues. Dans le cas particulier des articles scientifiques, Bennett/Muresan (2016 : 101) constatent une majorité de phrases contenant plus de 70 mots, dans un corpus portant sur le portugais. Fifielska (2015 : 25) compte une longueur moyenne de 28 à 29 mots par phrase pour le français, de 29 mots pour l'anglais et de 28,5 mots pour le russe. Hoffmann (1998 : 417) indique une moyenne de 15,9 mots par phrase pour les phrases simples (syntagmes avec un verbe conjugué au maximum) et de 33,5 mots par phrase pour les phrases complexes (coordination ou subordination d'au moins deux syntagmes verbaux contenant des verbes conjugués) dans le cas de l'allemand. Kocourek (1991 : 73) relève une longueur moyenne de 28,6 mots par phrase dans les textes technoscientifiques français.

Variation intra- et interdisciplinaire – Même si la tendance est à des phrases généralement plutôt longues dans l'écrit scientifique, il est important de considérer la « division horizontale » (Kocourek 1991 : 34, Reutner 2013 : 443*sq*.) du langage scientifique. Dans le cas de la phrase médicale par exemple, Rouleau (2006 : 140) relève une longueur moyenne de 24,6 mots par phrase, contre 23,8 mots dans son corpus de « textes généraux ». Or, cette différence s'avère statistiquement non significative, raison pour laquelle l'hypothèse selon laquelle « la phrase

technoscientifique se caractérise par sa longueur » (Kocourek 1991 : 73) peut être remise en question dans cette discipline. D'autres différences considérables entre les disciplines ont été présentées par Rinck (2006 : 189), qui relève une longueur moyenne de 19 mots par phrase dans les sciences du langage et de 23,5 mots par phrase dans le domaine des études littéraires. Elle en déduit une plus forte tendance à la structuration du texte dans les sciences du langage (*cf.* Rinck 2006 : 189). De plus, elle constate que l'écart type est plus faible dans les études littéraires (*cf.* Rinck 2006 : 190), la longueur n'y étant jamais inférieure à 15,3 mots par phrase, alors qu'elle varie entre 5 et 15 dans les sciences du langage. Elle en déduit une plus forte variation de la longueur des phrases dans cette dernière discipline (*cf.* Rinck 2006 : 190).

Changement diachronique – La longueur des phrases évolue au fil du temps et peut être un indice (parmi d'autres) d'un changement de style considérable. Dans le cas de l'anglais, l'étude de Rudnicka (2018 : 233) observe un recul considérable de la longueur des phrases dans les magazines, de 27,29 mots par phrase en 1810 à 17,14 mots par phrase en 2000. Ce résultat s'observe aussi en l'espace de 30 ans, comme le montre l'étude de Fengxiang (2007 : 130), qui indique une longueur moyenne de 21,16 mots par phrase dans le *Corpus d'Oslo/Bergen* (LOB) des années 1950/1960, contre 19,68 mots par phrase dans le BNCA et 19,44 dans le BNCB, ces deux sous-corpus provenant du *British National Corpus* (BNC) des années 1980/1990. La tendance est valable pour toute une gamme des traditions discursives comme les magazines, les journaux, la fiction et les textes scientifiques (*cf.* Rudnicka 2018 : 232). Cette tendance se montre par ailleurs dans les textes scientifiques aussi bien anglais que français et allemand (*cf.* Rudnicka 2018 : 222). Les études présentent également plusieurs explications par rapport à ce changement de style. Rudnicka (2018 : 224) explique la tendance en évoquant un changement des conventions de ponctuation marquées par le recul du point-virgule. D'autres études invoquent un style plus dense et moins explicite, qui serait par exemple lié à l'usage de syntagmes prépositionnels abstraits remplaçant des relatives (p. ex. *experiments in India* au lieu de *experiments that were conducted in India* ; *cf.* Gray/Biber 2018 : 142) ou encore au remplacement du connecteur *in order to + infinitif* par le simple *to + infinitif* (*cf.* Rudnicka 2018 : 236).

3. Méthode

Dans cette section méthodologique, nous décrivons notre corpus (*cf.* 3.1), la méthode mise en place pour déterminer la longueur des phrases du corpus (*cf.* 3.2) ainsi que notre protocole d'analyse statistique (*cf.* 3.3).

3.1 Corpus

Revues – Étudiant principalement les différences de longueurs de phrases dans quatre disciplines scientifiques, nous avons tout d'abord procédé à un choix d'articles en tentant de réduire au maximum la variation de tout autre facteur pouvant également avoir une influence sur cette longueur. Nous avons donc veillé à ne prendre en considération que des articles provenant d'une seule revue par discipline. Pour ce qui est du domaine des sciences naturelles tout d'abord, le choix s'est porté sur la revue de biochimie médicale et clinique *Transfusion clinique et biologique*. Ce choix n'est pas fortuit. Il est tout d'abord dû à des restrictions d'ordre pratique : il n'existe en effet que peu de disciplines au sein des sciences naturelles utilisant actuellement encore des langues autres que l'anglais, la plupart des revues francophones ayant cessé leurs activités de publication (*cf.* Larivière 2019 : 13). La biochimie médicale et clinique fait partie des quelques exceptions à cette tendance, l'utilisation du français, même si minoritaire, y étant relativement régulière. Par ailleurs, le choix porté sur cette revue en particulier a été effectué sur la base des résultats de l'indicateur d'influence scientifique du *Scimago Journal Ranking* (https://www.scimagojr.com/ ; 20.12.2020), l'un des classements internationaux des revues les plus couramment consultés. Ce faisant, nous avons pris le journal francophone catégorisé sous *biochemistry* le mieux classé par cet indicateur.

Dans le domaine de la phonétique plus encore qu'en biochimie médicale et clinique, il n'existe actuellement plus de revue qui contiendrait un nombre suffisant d'articles publiés en français pour justifier sa prise en considération dans le présent corpus. C'est la raison pour laquelle notre choix s'est porté sur le recueil d'articles des *Journées d'Étude sur la Parole* 2018 de l'*Association Francophone de la Communication Parlée*, une sous-association francophone de l'*International Speech Communication Association*. Ce choix pourrait certes être problématique, étant donné que le texte peut être dans le cas de tels actes de congrès, selon l'auteur·e, plus ou moins proche de la communication orale qui le précède, ce qui peut poser des problèmes de comparabilité avec les articles de revues. Étant donnée l'absence d'alternative, ce choix était néanmoins difficilement contournable.

Pour ce qui est des domaines de la sociolinguistique et de la littérature, le problème de la langue ne s'est pas posé étant donnée l'existence de diverses revues publiées uniquement en français. Dans le cas de la sociolinguistique, notre choix s'est porté sur la revue *Langage et société*, revue comportant essentiellement des articles en français portant sur le langage, les langues et les discours en tant que phénomènes sociaux. Dans le cas des études littéraires finalement, c'est sur la revue *Études théâtrales* que nous avons porté notre choix, une revue de réflexion sur le fait théâtral combinant diverses approches (p. ex. dramaturgiques, littéraires ou encore esthétiques).

Auteur·e·s – Notons que les traditions par rapport au nombre d'auteur·e·s impliqué·e·s dans les articles sont très différentes d'une discipline à l'autre : alors qu'en biochimie médicale et clinique leur nombre varie entre 1 et 20 ($\overline{x}$ = 4,90 ; σ = 4,88) et en phonétique de 1 à 5 ($\overline{x}$ = 3,55 ; σ = 1,36), tous les articles de sociolinguistique et de littérature ont été écrits par un·e seul·e auteur·e. Ces différentes traditions de rédaction peuvent donc également avoir eu un effet

considérable sur le langage utilisé des articles (et sur la longueur des phrases) et devront donc être prises en compte dans l'interprétation des données.

Articles – Au sein de ces quatre revues, nous avons uniquement pris en considération des articles d'auteur·e·s francophones affilié·e·s à des universités françaises. Ce choix a été fait afin de réduire au mieux le facteur de la socialisation académique des auteur·e·s, qui peut fortement influer sur les données. Par ailleurs, nous avons uniquement pris en considération les articles les plus récents, débutant ainsi la recherche par les numéros les plus récents et allant ensuite dans un ordre décroissant de parution (*Transfusion clinique et biologique* : 2016–2019 ; *Journées d'Étude sur la Parole* : 2018 ; *Études théâtrales* : 2013–2014 ; *Langage et société* : 2014–2016). Notons également que nous n'avons pris en compte que des articles scientifiques originaux, tout autre type d'article (p. ex. les comptes rendus, les éditoriaux ou encore les cas cliniques et les perspectives de recherche dans le cas du journal de biochimie médicale et clinique) ayant été retiré du corpus avant son analyse. De cette manière, 20 articles ont été tirés dans chacune des quatre revues pour un total de 80 articles.

Sections – Notre analyse porte sur les 10 premières phrases de la *conclusion* de chacun des articles, pour un total de 800 phrases dans le corpus. Cette section a été choisie étant donné qu'il s'agit souvent de l'une des sections conçues particulièrement consciemment étant donné que les principaux résultats ainsi que leurs implications y sont présentés et que des questions ouvertes appellent le lecteur à des recherches ultérieures. Cependant, ce choix pourrait poser deux problèmes méthodologiques qu'il est important de mentionner. Premièrement, tous les articles ne comportent pas forcément de section intitulée *conclusion*. C'est notamment le cas de la plupart des articles de la revue *Études théâtrales*, au sein de laquelle les sections sont en général thématiques et ne suivent pas le schéma de l'IMRAD (Introduction, Méthode, Résultats et (and) Discussion) couramment utilisé en particulier dans les sciences naturelles (*cf.* Pontille 2003 : 56). Dans ces cas, ce sont les dix dernières phrases des articles qui ont été prises en considération. En définitive, cette différence de macrostructure des articles ne s'avère qu'en partie problématique étant donné que nos analyses ont pu montrer qu'implicitement, les dernières phrases des articles sans conclusion explicite remplissent également une fonction conclusive, ce qui est illustré par les deux exemples suivants : « En cela, Robespierre forme une pièce de fin de vie, qui conclut de manière parfaitement cohérente un cycle dramatique commencé à l'orée de la carrière littéraire de son auteur » (corpus littérature_1) ; « Cet appel d'air qui nous projette dans un extérieur inconnu […] serait celui de la mort, si la musique, c'est-à-dire le son, cessait. Il persiste. » (corpus littérature_6). Ce type de phrases conclusives apparaissant avec une certaine régularité, notamment dans la revue *Études théâtrales*, le manque de comparabilité apparent peut être – du moins en partie – relativisé. Deuxièmement, certaines conclusions ne comportaient pas les 10 phrases requises. Dans ces cas, les phrases prises en compte ont été complétées par des phrases de la section précédant directement la conclusion, ce qui sera pris en compte dans l'interprétation des données.

3.2 Identification de la longueur des phrases

Passons maintenant au trait retenu pour analyser les différences de langage dans les quatre disciplines. Notre analyse se base sur la *longueur* des dix phrases prises en compte dans chacun des articles. Il ne s'agit bien sûr que de l'un des nombreux critères pouvant être retenus pour analyser de telles différences, à côté de la complexité syntaxique, de la structure thème-rhème, de la condensation syntaxique, de l'impersonnalité, de la nominalisation des prédicats ou encore de la désémantisation des verbes (*cf.* p. ex. Simmler 2006 : 1523 ; Petkova-Kessanlis 2015 : 212). Par ailleurs, il faut garder à l'esprit qu'au contraire de ces derniers critères, qui sont syntaxiques, la longueur des phrases telle que nous l'analysons – sur la base de la majuscule et du point – est un critère plus graphique que syntaxique (*cf.* 2 : *Caractéristiques statistiques*).

Le choix de ce critère nous semble cependant central lorsqu'il s'agit de déterminer une différence entre un style de rédaction plutôt bref et un style plutôt diffus, comme le suggère le vieux stéréotype caractérisant les sciences naturelles comme étant 'avares en mots' et les sciences humaines comme présentant un style diffus 'en étalant trop'. Trois autres caractéristiques de ce trait justifient également notre choix. Premièrement, il est idéal dans le cadre de l'analyse quantitative d'un corpus étant donné qu'il est déterminable d'une manière tout à fait précise. Deuxièmement, des études publiées précédemment ont pu montrer que la longueur des phrases fait partie des traits permettant particulièrement bien de distinguer les différentes traditions discursives (*cf.* p. ex. Karlgren/Cutting 1994 : 1073), l'écrit scientifique ayant par exemple tendance – même si cette tendance est actuellement en baisse – à présenter des phrases relativement longues par rapport à d'autres types de textes (*cf.* Kelih *et al.* 2006 : 386). Troisièmement, il s'agit d'un trait qui était à la base des études montrant en particulier des différences interdisciplinaires entre sciences du langage et études littéraires (*cf.* p. ex. Rinck 2006 : 189). L'utilisation de ce même trait permet donc une meilleure comparabilité de notre étude avec les études précédentes. Mais il est évident qu'il ne peut à lui seul expliquer qu'une infime partie des différences entre les pratiques de rédaction dans les disciplines et que les hypothèses qui sont tirées de nos analyses devront être vérifiées à l'avenir sur la base d'autres traits mentionnés ici.

Notre méthode de comptage repose sur les principes suivants. Premièrement, elle se base sur les mots graphiques, ce qui comporte certains aspects délicats qui doivent être mentionnés. C'est notamment le cas des composés lexicaux détachés (p. ex. *hors de propos*), souvent difficiles à distinguer de simples groupes de mots. Dans notre corpus, aucune différence n'est faite entre ces deux catégories étant donné que chaque mot graphique est compté. Le problème ne se pose, au contraire, pas dans le cas des composés lexicaux unifiés (p. ex. *lequel*), à traits d'union (p. ex. *celui-ci*) et à apostrophe (p. ex. *aujourd'hui*), qui ne comptent respectivement que comme un seul et unique mot. Cependant, cette légère distorsion des résultats aura tendance à être la même dans les quatre disciplines, de sorte que la comparaison de ces disciplines n'en est que peu altérée. Deuxièmement, notons qu'en conformité avec les études quantitatives de Best

(p. ex. 2005 : 300), les points-virgules (« ; ») et les deux-points (« : ») n'ont pas été considérés comme des frontières de phrases. Cette décision peut également être perçue comme problématique étant donné que, selon l'auteur·e, des phrases syntaxiquement complètes peuvent suivre les points-virgules et les deux-points. Ici aussi, la potentielle distorsion des résultats sera cependant similaire dans les quatre disciplines, de sorte que la tendance globale pourra certes être légèrement altérée, mais la comparaison des disciplines entre elles en restera tout de même valide. Notons ce faisant que l'utilisation des points-virgules et des deux-points sera tout de même quantifiée dans les données et prise en compte dans l'interprétation des résultats. Troisièmement, certains articles comportaient des sous-titres dans la conclusion, qui n'ont pas été pris en considération dans le comptage des mots. De la même manière, toute phrase comprenant de longues citations directes de plus de cinq mots a été supprimée du corpus et remplacée par la phrase suivante. Quatrièmement, les références citées entre parenthèses ont été prises en compte dans le calcul, de même que les indications de pourcentages ou autres chiffres (p. ex. « 36% »), qui ont été respectivement considérés comme équivalant à un mot. Notons finalement que le comptage des mots a été effectué manuellement afin d'éviter toute erreur de calcul pouvant apparaître dans les comptages automatiques effectués avec des logiciels de reconnaissance optique de caractères (OCR ; *cf.* Niemann 2018 : 109).

3.3 Protocole d'analyse

L'exploitation statistique des données ainsi récoltées s'est faite en deux étapes : elles ont été structurées sous *Excel* et soumises à des analyses statistiques à l'aide du logiciel libre *R*. Ce faisant, elles ont été soumises à une analyse sur la base d'un modèle à effets mixtes (fonction *lmer* de l'extension *lme4*) combinée à un test de Tukey (fonction *glht* de l'extension *mutlcomp*). Statistiquement, ce choix se justifie par le fait que notre corpus contient une variable explicative aléatoire (les 80 articles analysés), une variable explicative fixe (le regroupement des articles par 20 dans quatre disciplines différentes : biochimie médicale et clinique, phonétique, sociolinguistique, littérature) et la variable dépendante de la longueur des phrases.

Notons ce faisant qu'aucune autre variable explicative n'a été ajoutée au modèle, la variable de l'âge, souvent considérée comme particulièrement pertinente dans l'étude de productions écrites, n'y faisant pas exception. Ce dernier choix se justifie par le fait que les articles pris en compte n'ont été rédigés par un·e seul·e auteur·e qu'en sociolinguistique et dans les études littéraires, alors qu'en biochimie médicale et clinique et en phonétique, la règle est aux auteur·e·s multiples (*cf.* 3.1 - *Auteur·e·s*). Par ailleurs, tout article scientifique est systématiquement retravaillé par un comité de lecture, notamment dans les revues fonctionnant sur la base d'évaluations par les pairs (*peer-review*), l'influence de ces pairs pouvant même toucher les choix linguistiques de l'auteur·e.

Soulignons en outre que les résultats statistiques ont été complétés par des analyses qualitatives ponctuelles permettant d'expliquer certaines tendances (*cf.* 4.3). Finalement, notons que dans la section 4.2, nous avons complété nos résultats sur les schémas pouvant expliquer la longueur des phrases par une analyse des *éléments péritextuels*[2] présents dans les 80 textes ainsi que de la longueur des phrases les introduisant. Dans ce sous-corpus, nous ne nous sommes, au contraire du cas des autres analyses, pas restreints aux conclusions des articles, mais avons pris en compte la globalité de chacun des 80 articles.

4. Résultats et discussion

Notre corpus a tout d'abord été analysé sur la base d'un modèle à effets mixtes. La longueur des phrases a ce faisant été définie comme variable dépendante, les quatre disciplines comme variable explicative à effets fixes et les 80 articles du corpus comme variable à effets aléatoires (*cf.* tableau 1).

Tableau 1. Modèle à effets mixtes appliqué à la longueur des phrases dans le corpus (fonction *lmer* de l'extension *lme4*).

	Variable dépendante			
Facteurs	Longueur des phrases			
	Coefficient	Erreur standard	Valeur *t*	Valeur *p*
Effets fixes				
Disciplines	29,52	1,81	16,27	<0,001
	Variance	Écart type		
Effets aléatoires				
Articles	42,81	6,54		
Observations	800			
Articles	80			

Le modèle présente principalement deux résultats distincts. Premièrement, sur la base d'une variance (42,81) et d'un écart type (6,54) plutôt élevés, il montre de forts effets aléatoires des 80 articles sur la longueur des phrases. En d'autres mots et sans grande surprise, l'influence des auteur·e·s sur la longueur des phrases est donc considérable.

Deuxièmement, mis à part ces effets aléatoires, le facteur de la discipline présente également une influence sur les longueurs de phrases mesurées, cette influence s'avérant même hautement significative à $p<0{,}001$. La présentation des résultats détaillés sera faite sur la base de cette bipartition entre effets fixes (*cf.* 4.1) et effets aléatoires (*cf.* 4.2). La variation des longueurs de phrases observée sur la base des effets aléatoires sera par ailleurs complétée par une analyse qualitative des schémas observables dans l'alternance de la longueur des phrases (*cf.* 4.3).

4.1 Influence de la discipline de spécialisation sur la longueur des phrases (effets fixes)

En ce qui concerne les effets fixes des quatre disciplines sur la longueur des phrases mesurées, une visualisation des données sous forme de diagramme en boîte illustre bien les différences notées par le modèle (*cf.* fig. 1).

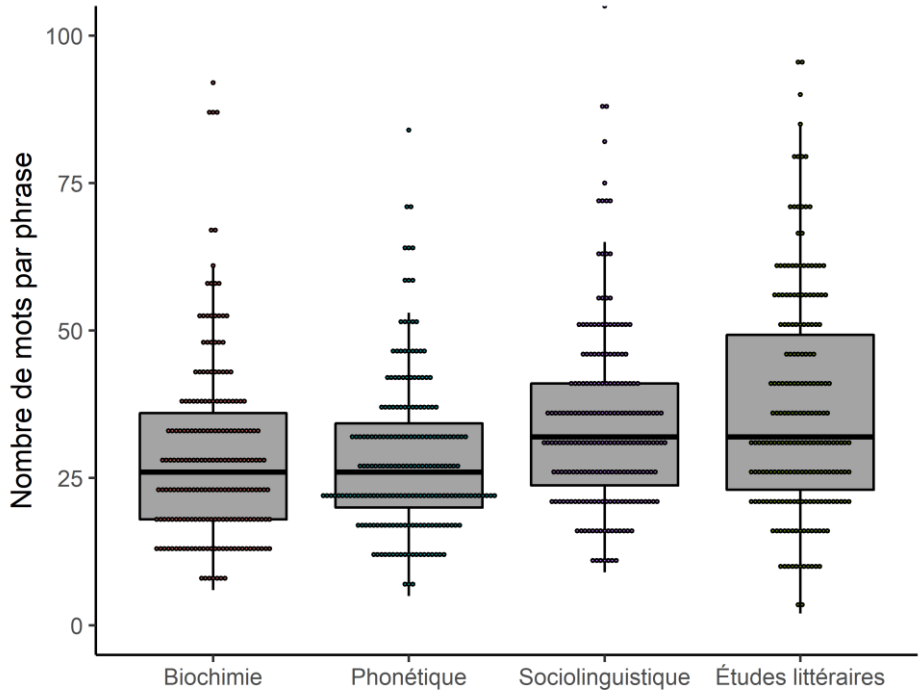


Fig. 1. Comparaison de la longueur des phrases dans les articles de biochimie médicale et clinique, de phonétique, de sociolinguistique et d'études littéraires ($p<0{,}001$, coef. = 29,52 ; e.s. = 1,81 ; $t = 16{,}27$).

Le diagramme montre pour chacune des disciplines la médiane (lignes horizontales à l'intérieur des boîtes), l'écart interquartile qui couvre 50% des données de chacun de ces groupes et qui élimine les valeurs aberrantes (boîtes grises), l'étendue de ces données (lignes verticales centrales) ainsi que le degré de fluctuation des données sur cette étendue (pointillés le long de l'axe y). Globalement, une différence est observable entre les médianes mesurées, d'une part, dans le cas de la biochimie médicale et clinique et de la phonétique (respectivement 26 mots) et, d'autre part, dans les cas de la sociolinguistique et des études littéraires (respectivement 32 mots). Par ailleurs, la dispersion des données des études littéraires, visible sur la base d'une étendue et d'un écart interquartile plus élevés, semble largement plus élevée que celle des trois autres disciplines. Le modèle à effets mixtes (*cf.* tableau 1) confirme globalement l'existence de telles différences à $p<0.001$, mais sans indiquer entre quels groupes les différences sont significatives et entre lesquels elles ne le sont pas. C'est la raison pour laquelle nous avons également procédé à une comparaison de chacun de ces groupes avec les trois autres groupes dans le cadre d'un test de Tukey. Les tendances de ce test sont résumées dans le tableau 2.

Tableau 2. Comparaison de chaque discipline avec les trois autres par rapport à la longueur des phrases sur la base d'un test de Tukey (fonction *glht* de l'extension *mutlcomp* sous *R*).

Paires comparées		Diff.	Valeur *z*	Valeur *p*
Biochimie	Phonétique	0,96	0,37	>0,05
	Sociolinguistique	-4,83	-1,88	>0,05
	Études littéraires	-7,76	-3,02	<0,05
Phonétique	Biochimie	-0,96	-0,37	>0,05

	Sociolinguistique	-5,79	-2,26	>0,05
	Études littéraires	-8,71	-3,40	<0,01
Sociolinguistique	Biochimie	4,83	1,88	>0,05
	Phonétique	5,76	2,26	>0,05
	Études littéraires	-2,93	-1,14	>0,05
Études littéraires	Biochimie	7,76	3,02	<0,05
	Phonétique	8,71	3,40	<0,01
	Sociolinguistique	2,93	1,14	>0,05

Comme le montrent les tendances illustrées dans la figure 1 et les taux de signification présentés dans le tableau 2, les données de notre corpus semblent confirmer en partie notre principale hypothèse. Les valeurs *p* révèlent en effet une hiérarchie dans les disciplines quant à la longueur moyenne des phrases utilisées. La biochimie médicale et clinique et la phonétique, qui présentent les phrases les plus courtes du corpus avec des médianes de 26 mots et des écarts interquartiles similaires (biochimie : 18 ; phonétique : 14,25 ; *cf.* fig. 1), ne se différencient pas significativement l'une de l'autre ($p>0,05$). À l'autre extrémité, les études littéraires présentent pour leur part les longueurs de phrase les plus longues du corpus (médiane : 32 mots ; écart interquartile : 26,25 ; *cf.* fig. 1), ces valeurs se différenciant significativement de celles de la biochimie médicale et clinique ($p<0,05$) et de la phonétique ($p<0,01$). La sociolinguistique, finalement, montre des valeurs se trouvant entre ces deux extrémités (médiane : 32 mots ; écart interquartile : 17,25 ; *cf.* fig. 1) et ne se différenciant ni de celles de la biochimie médicale et clinique ou de la phonétique, ni de celles des études littéraires ($p>0,05$). Ainsi, les différences qui avaient été mises en lumière notamment par Rinck (2006) entre les sciences du langage et les études littéraires (*cf.* 2 : *Variation intra- et interdisciplinaire*) sont globalement confirmées par nos résultats. Ces derniers permettent cependant, étant donnée la distinction supplémentaire faite entre la phonétique et la sociolinguistique, d'ajouter une nuance supplémentaire interne aux sciences du langage : la phonétique présente en effet un comportement plus proche de celui des sciences naturelles, alors que la sociolinguistique est plus proche de celui des études littéraires.

Deux observations se doivent maintenant d'être faites par rapport à ce résultat global. La première concerne l'influence des phrases contenant des points-virgules et des deux-points sur la longueur des phrases mesurées, étant donné que nous n'avons pas considéré ces signes de ponctuation comme frontières de phrases. La deuxième concerne le résultat plutôt surprenant, bien que non significatif, montrant des phrases globalement plus courtes en phonétique qu'en biochimie. Abordons tout d'abord l'utilisation des points-virgules et des deux-points. Notons qu'une quantification de ces signes de ponctuation mène à des tendances très similaires de celles touchant aux longueurs de phrases. Ils sont en effet utilisés considérablement plus souvent dans les études littéraires et en sociolinguistique qu'en biochimie médicale et clinique et qu'en phonétique, comme le montre la figure 2.

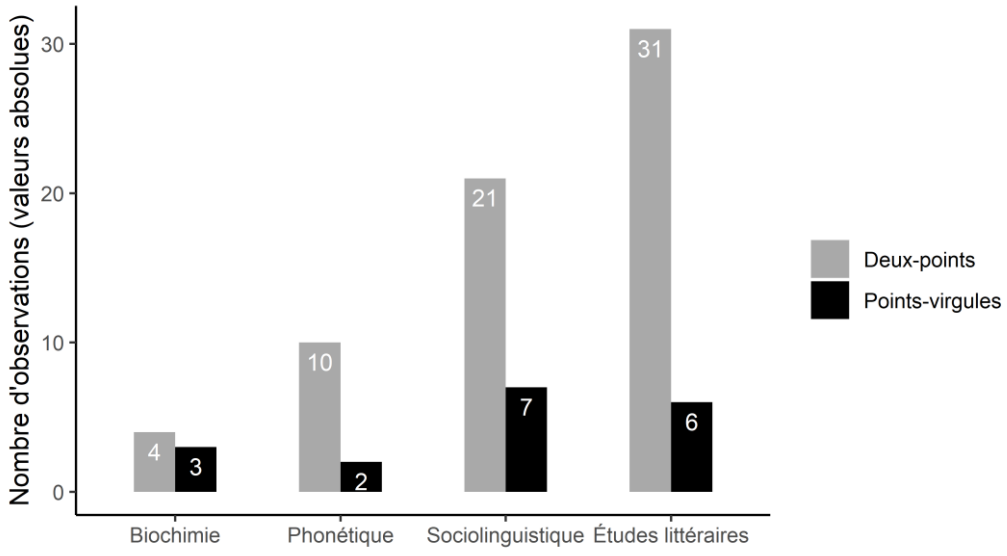


Fig. 2. Comparaison de l'utilisation des deux-points et des points virgules dans les articles de biochimie médicale et clinique, de phonétique, de sociolinguistique et d'études littéraires (*ANOVA* : $F(3,796) = 9,42$; $p<0,001$).

Si la tendance observée dans le cas des points-virgules n'est statistiquement pas exploitable étant donné le trop faible nombre de données (18 observations au total), la tendance à l'utilisation plus ou moins fréquente des deux-points présente globalement des différences significatives à $p<0,001$ ($F(3,796) = 9,42$). Un test de Tukey révèle que l'utilisation des deux-points en biochimie médicale et clinique et en phonétique est significativement différente de celle observée dans les études littéraires ($p<0,001$). Par ailleurs, ces tests montrent également le statut intermédiaire de la

sociolinguistique étant donné que l'utilisation des deux-points en sociolinguistique se différencie significativement de celle de la biochimie médicale et clinique (p<0,05), mais pas de celle de la phonétique (p>0,05) (*cf.* tableau 3).

Tableau 3. Comparaison des quatre disciplines entre elles par rapport à l'utilisation des deux-points sur la base d'un test de Tukey (fonction *glht* de l'extension *mutlcomp* sous *R*).

Paires comparées		Diff.	Valeur *z*	Valeur *p*
Biochimie	Phonétique	-0,03	-1,09	>0,05
	Sociolinguistique	-0,09	-3,09	<0,05
	Études littéraires	-0,14	-3,81	<0,001
Phonétique	Biochimie	0,03	1,09	>0,05
	Sociolinguistique	-0,06	-1,97	>0,05
	Études littéraires	-0,11	-3,81	<0,001
Sociolinguistique	Biochimie	0,09	3,09	<0,05
	Phonétique	0,06	1,97	>0,05
	Études littéraires	-0,05	-1,82	>0,05
Études littéraires	Biochimie	0,14	-3,81	<0,001
	Phonétique	0,11	3,81	<0,001
	Sociolinguistique	0,05	1,82	>0,05

Les tendances étant ainsi très similaires à celles observées pour la longueur des phrases (*cf.* tableau 2), les phrases contenant des points virgules et des deux-points auront effectivement pu exercer une influence sur le résultat global de la longueur des phrases. Cependant, au total, seul un dixième des phrases (84/800) est touché par cette particularité, de sorte que si influence il y a, celle-ci reste relativement faible.

Passons maintenant à la deuxième observation, qui montre de manière plutôt surprenante, malgré le manque de signification statistique de cette différence, des phrases globalement plus courtes en phonétique qu'en biochimie médicale et clinique. Une analyse plus détaillée des données fournit cependant une explication plausible à ce résultat. Les 20 articles de biochimie médicale et clinique comprennent en effet deux types d'articles : des articles de biochimie expérimentale, d'une part, et des articles liés à certaines questions d'éthique en biochimie, d'autre part. Or, il se trouve que les longueurs de phrases se différencient significativement l'une de l'autre à p<0,05 dans ces deux catégories, avec des phrases plus courtes dans le domaine de la biochimie expérimentale ($\overline{x}$ = 28,15 ; σ = 16,60) et des phrases plus longues dans le domaine de l'éthique en biochimie ($\overline{x}$ = 33,60 ; σ = 14,86) (*cf.* fig. 3).

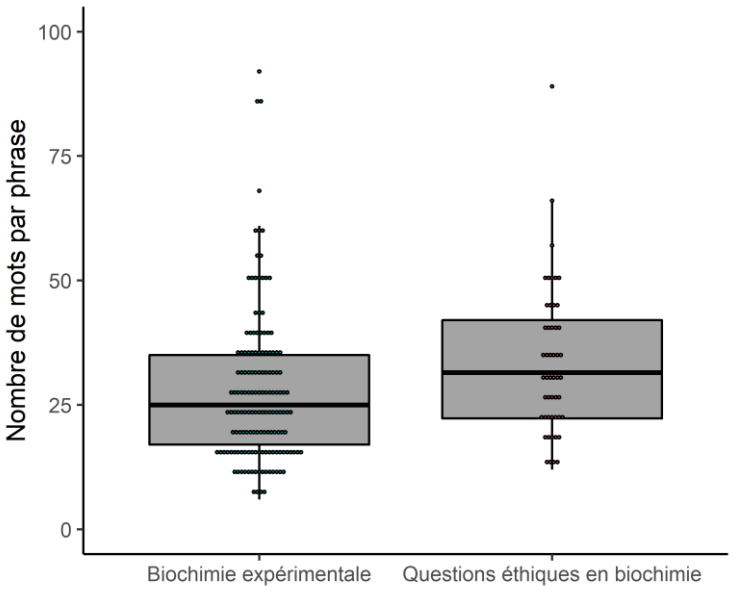


Fig. 3. Comparaison de la longueur des phrases dans les articles de biochimie expérimentale et de biochimie éthique (*ANOVA* : $F(1,198) = 12,44$, p<0,05).

Ainsi, en plus de dépendre de la tradition de rédaction des disciplines (*cf.* tableau 2), la longueur des phrases semble également dépendre de la méthode de travail employée. Les méthodes empiriques, qui se basent généralement sur des expériences, des observations ou des sondages, semblent s'avérer plus propices aux phrases courtes. Les méthodes non-empiriques dans lesquelles les résultats peuvent être présentés sans recours à l'observation directe, s'avèrent propices à l'utilisation de phrases plus longues. Nous présentons à titre illustratif des phrases prototypiques illustrant ces observations dans le tableau 4, en étant bien conscients que les quatre domaines comprennent des phrases des deux types.

Tableau 4. Exemples prototypiques illustrant la tendance aux phrases courtes ou longues selon la méthode de travail de l'auteur·e.

Observations empiriques directes : Phrases plutôt courtes	
Biochimie	« La transfusion sanguine est un acte thérapeutique fréquent en néonatologie. » (10 mots ; corpus biochimie_3)
Phonétique	« Seule la correction des erreurs de classification fait diminuer directement le DER. » (12 mots ; corpus phonétique_9)
Processus d'introspection ou de réflexion : Phrases plutôt longues	
Sociolinguistique	« La sociolinguistique de terrain se distingue donc d'une linguistique où le terrain ne serait qu'un décor, tout comme d'une linguistique qui détache les éléments langagiers de leurs conditions de production et de réception et, partant, d'une linguistique qui prétend neutraliser ou contrôler des paramètres afin de vérifier des hypothèses construites compte non tenu des réalités mouvantes et complexes des échanges sociaux. » (65 mots ; corpus sociolinguistique_10)
Études littéraires	« À travers l'oscillation entre des postures incarnées et désincarnées, oscillation qui opère au croisement des interactions entre marionnette et être humain, voix en présence et voix en play-back, mais surtout entre ce que le texte dit et ce que la voix et le corps révèlent, se joue le combat entre une instance informe, irreprésentable, et le sujet, figé dans des masques et dans une langue de bois. » (68 mots ; corpus littérature_2)

Ainsi, le schéma suivant est relativement frappant dans ces données : les *faits empiriquement observables* semblent souvent être décrits à l'aide de phrases plutôt courtes. Au contraire, les *processus d'introspection ou de réflexion*, souvent considérés comme typiques les sciences humaines (*cf.* p. ex. Paillé/Mucchielli 2012 : 103), semblent favoriser l'utilisation de phrases plus longues. Cette hypothèse confirmerait les observations de Barr (2001 : 378), qui avait déjà fait le lien entre réflexions et phrases sensiblement plus longues, ce qu'il explique en invoquant ce qu'il nomme le « flux de pensées ininterrompu » (*cf.* Barr 2001 : 378).

Bien que la présente étude ne se focalise pas sur cet aspect, il est intéressant de constater que les éléments péritextuels – parmi lesquels nous ne prenons ici, à titre exemplaire, que les tableaux, les graphiques, les images et illustrations en compte – ainsi que les phrases introduisant ces éléments confirment globalement ce schéma : il est tout d'abord observable que les domaines de la biochimie (2,40/article en moyenne) et de la phonétique (4,15/article) utilisent largement plus d'éléments péritextuels que les domaines de la sociolinguistique (1,40/article) et des études littéraires (0,05/article) (*cf.* tableau 5). Par ailleurs, ces éléments se différencient également par leur nature : une grande majorité des tableaux et graphiques trouvés dans les domaines de la biochimie (87,50%) et de la phonétique (75,90%) consistent en des tableaux et des schémas visualisant ou modélisant des données statistiques (*faits empiriquement observables*), alors que dans les domaines de la sociolinguistique et des études littéraires, nous avons plus à faire à des schémas interprétatifs ou classificatifs qu'à des statistiques (*processus de réflexion*), le taux d'éléments péritextuels statistiques s'élevant dans ces deux dernières disciplines à 39,29% (sociolinguistique) et 0,00% (études littéraires) (*cf.* tableau 5). Ainsi, les *faits empiriquement observables* ne semblent donc pas seulement favoriser les phrases courtes, mais également l'utilisation d'éléments péritextuels, en majorité statistiques. Les processus d'introspection ou de réflexion, pour leur part, sont exprimés à l'aide de phrases plutôt longues et en général uniquement sous forme de texte. Finalement, notons aussi qu'il existe une corrélation entre les éléments péritextuels et les phrases courtes : en effet, dans notre corpus, la phrase introduisant explicitement un tableau ou graphique[3] est en moyenne significativement plus courte que la longueur moyenne des phrases mesurées dans le corpus global. Cette différence est valable à $p<0,05$ (selon un test de Student) pour trois des quatre domaines scientifiques (biochimie, phonétique, sociolinguistique), les études littéraires y faisant exception étant donné qu'un seul élément péritextuel a pu y être trouvé et qu'aucune mesure statistique n'a donc pu être effectuée dans ce domaine (*cf.* tableau 5).

Tableau 5. Éléments péritextuels et leur corrélation avec les phrases courtes les introduisant

Discipline	Éléments péritextuels ($\bar{x}$)	Éléments péritextuels statistiques[4] (%)	Longueur des phrases introduisant un élément péritextuel ($\bar{x}$)	Longueur des phrases dans le corpus global ($\bar{x}$)
Biochimie	2,40 (48/20)	87,50% (42/48)	22,71	29,52
Phonétique	4,15 (83/20)	75,90% (63/83)	26,56	28,56
Sociolinguistique	1,40 (28/20)	39,29% (11/28)	29,38	34,35
Études littéraires	0,05 (1/20)	0,00% (0/1)	∅[5]	37,27

Rappelons ici qu'il ne s'agit que de tendances observées sur la base de données fortement variables (*cf.* tableau 1 ; effets aléatoires). Notre hypothèse (*faits empiriquement observables* : phrases plutôt courtes combinées à de nombreux éléments péritextuels, en général statistiques ; *processus d'introspection ou de réflexion* : phrases plutôt longues et faible nombre d'éléments péritextuels), qui est inévitablement momentanée et réductrice, se doit donc d'être interprétée avec précaution. Ceci est d'autant plus important que plusieurs études ont dans un passé proche déjà pu observer un recul global considérable de la longueur des phrases (*cf.* p. ex. Rudnicka 2018 : 222), même si – rappelons-le – ce recul affecte toutes les disciplines de manière relativement égale.

4.2 Forte variation des longueurs de phrases parmi les 80 articles (effets aléatoires)

En ce qui concerne la variabilité des longueurs de phrases, mentionnons tout d'abord que les écarts types relevés dans les données de chacune des disciplines montrent une forte dispersion des données autour des quatre moyennes, l'intervalle de fluctuation contenant la majorité des longueurs de phrases mesurées étant ainsi très large dans les quatre disciplines (*cf.* tableau 6).

Tableau 6. Moyennes, médianes et indices de dispersion des données dans les quatre disciplines.

Discipline	Moyenne	Médiane	Écart interquartile	Écart type	Intervalle de fluctuation
Biochimie	29,52	26	18,00	15,53	15,46 – 41,66
Phonétique	28,56	26	14,25	13,10	13,99 – 45,05
Sociolinguistique	34,35	32	17,25	16,32	18,03 – 50,67
Études littéraires	37,27	32	26,25	20,12	17,15 – 57,39

Cette forte dispersion des données peut tout d'abord être mise sur le compte de la variabilité des données entre les 80 articles, ce que montrent les forts effets aléatoires du modèle mixte utilisé (variance : 42,81 mots ; écart type : 6,54 mots). Par ailleurs, cette variabilité se retrouve même à l'intérieur même d'un article, ce qui n'a certes pas été mesuré statistiquement en raison du faible nombre de données analysé à l'intérieur de chaque article (10 phrases), mais qui est clairement visible dans la figure 4. Cette dernière présente un diagramme en boîte pour chacun des 80 articles, avec la valeur médiane en gras et la dispersion des données entre le premier (frontière inférieure de la boîte) et le troisième quartile (frontière supérieure de la boîte).

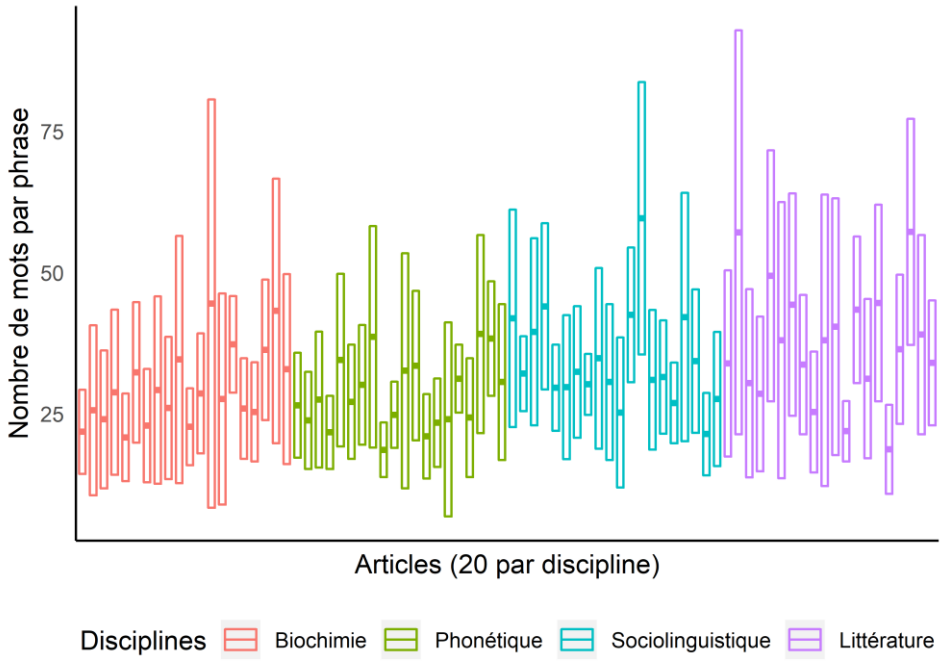


Fig. 4. Dispersion des données à l'intérieur de chacun des 80 articles analysés.

Retenons donc que malgré une tendance globale à la croissance de la longueur des phrases selon la hiérarchie biochimie médicale et clinique/phonétique > sociolinguistique > études littéraires, la variabilité entre les articles et à l'intérieur de chaque article est considérable. L'observation faite par Barr (2001 : 377) par rapport à la forte variabilité des longueurs de phrases dans le discours scientifique écrit s'avère donc largement confirmée dans notre corpus.

4.3 Alternance de la longueur des phrases : effets stylistiques

Maintenant, en plus de la variation aléatoire de la longueur des phrases qui semble inhérente à tout texte (*cf.* 2 : *Variation intra- et interdisciplinaire*), une analyse qualitative plus détaillée à l'intérieur de chacun des articles montre un autre aspect apportant une explication supplémentaire à cette variation. Il semble en effet exister une tendance générale à l'alternance régulière de phrases plutôt courtes et de phrases plutôt longues. Afin de mettre en lumière les fonctions de cette alternance de manière qualitative et à titre exemplaire, nous avons pris en considération toutes les phrases se trouvant en-dessous de la valeur minimale de l'intervalle de fluctuation (*cf.* tableau 6) et étant entourées par des phrases plus longues. Au total, notre corpus contient 39 de ces phrases courtes entourées de phrases plus longues. Or, ces phrases courtes peuvent revêtir trois fonctions : 19 (19/39) d'entre elles ont une valeur *introductive*, 9 (9/39) d'entre elles une valeur *conclusive* et 11 (11/39) d'entre elles une valeur *contrastive*.

Valeur introductive – Notre corpus contient 19 phrases courtes (19/39) annonçant soit le début d'un nouvel aspect thématique, soit un élément péritextuel (*cf.* 4.1). Au sein de cette catégorie, 10 des 19 phrases (10/19) se trouvent au début et les 9 autres (9/19) au milieu d'un paragraphe. Dans l'exemple ci-dessous, une phrase très courte (10 mots), située au début de la section *discussion* et évoquant le sujet principal de l'article, la transfusion sanguine, est suivie d'une phrase de longueur moyenne selon notre intervalle de fluctuation (20 mots), cette dernière agissant comme porteuse des informations plus détaillées :

> [Début de la section *discussion*] La transfusion sanguine est un acte thérapeutique fréquent en néonatologie. Cette enquête fournit un descriptif des modalités transfusionnelles dans des établissements de santé (ES) français avec maternité de niveau III (*cf.* corpus biochimie_3).

Cette valeur introductive peut par ailleurs être soulignée par la présence d'éléments cataphoriques renvoyant à des éléments à venir. De tels éléments ont pu être trouvés dans 6 des 19 phrases présentant une valeur introductive (6/19). La cataphore peut tout d'abord, comme dans le premier exemple, être exprimée par une locution adverbiale (*comme suit*) annonçant la phrase qui suit. Dans le deuxième exemple, c'est l'expression *deux explications* qui revêt la fonction cataphorique et qui est reprise dans les phrases suivantes par *la première* et *la seconde*, qui pour leur part prennent, dans la perspective inverse, une fonction anaphorique. Dans le troisième exemple, tiré du corpus sur les études littéraires, une question rhétorique renvoie à un nouvel aspect thématique à venir. Dans cet exemple, la première phrase, d'une longueur supérieure à l'intervalle de fluctuation (63 mots), contient une description des spécificités dramaturgiques de la pièce analysée. Ces réflexions sont par la suite interrompues par une question rhétorique très courte (9 mots) introduisant une nouvelle pensée, qui est poursuivie dans une troisième phrase d'une longueur à nouveau largement supérieure à l'intervalle de fluctuation (71 mots).

> Elle se calcule comme suit (*cf.* corpus phonetique_15).

> Deux explications potentielles à cette contradiction peuvent être formulées. La première serait que les signaux EMG collectés sur l'électrode du MENT soient en fait contaminés par les potentiels d'action des muscles abaisseurs adjacents, en particulier le DLI (phénomène de diaphonie). La seconde serait qu'il existe un mécanisme de coactivation des muscles élévateurs et abaisseurs de la lèvre pour garantir une meilleure stabilité du geste d'ouverture, dont on sait qu'il est crucial pour la production des plosives (*cf.* corpus phonetique_11).

> L'auteur-marionnettiste schwabien paraît improviser selon les réactions de son public, être à la fois auteur, metteur en scène et, via ses figures, comédien puisque ce sont ses doigts qui jouent et que ce sont, dirait-on, les spectateurs qui lui disent ce qu'ils veulent voir, ou bien le cas échéant, comme dans *Übergewicht*, *unwichtig*: *Unform*, voir recommencer sur un autre mode. Le spectacle marionnettique schwabien est-il pour autant irréaliste ? Il est plutôt une amorce, une possibilité, quelque chose d'entrevu : ce sera au spectateur d'assurer le raccord avec ce qui aurait pu lui arriver (l'agôn) et avec ce qu'il aurait pu en penser (la morale de l'histoire), ou plutôt, car il convient ici d'abandonner la phraséologie conventionnelle, le raccord avec sa propre réalité, ce qu'il ne pourra faire qu'après le tomber du rideau (corpus littérature_11).

Valeur conclusive – À côté de cette valeur introductive, les phrases courtes peuvent également indiquer, dans la perspective inverse, la fin d'un aspect thématique (9/39 extraits). Dans l'exemple suivant, la première phrase, de longueur moyenne selon l'intervalle de fluctuation (21 mots), présente certains résultats de l'étude. La seconde phrase (11 mots), plutôt courte, pose pour sa part le champ d'application de ces résultats et marque de la même manière la fin

de l'aspect thématique. La troisième phrase, plutôt longue (34 mots), ajoute un aspect supplémentaire introduit par la locution adverbiale *de plus* permettant de continuer l'argumentation :

> De la même manière, la durée de la voyelle /a/ qui précède C2# est plus longue que lorsque suivie par C2σ. Ce résultat est valable pour les plosives comme pour les nasales. De plus, la durée d'une syllabe dépend du fait qu'elle constitue un mot monosyllabique […] (*cf.* corpus phonétique_13).

Valeur contrastive – Il est par ailleurs observable que les différentes longueurs peuvent également être utilisées pour souligner un contraste entre les phrases (11/39 extraits). Dans le premier exemple ci-dessous, la deuxième phrase, très courte (8 mots), souligne l'importance de recherches futures sur la base des informations données dans les phrases qui l'entourent, qui sont, avec respectivement 49 mots et 36 mots, plutôt longues selon l'intervalle de fluctuation. Le connecteur *néanmoins*, utilisé dans la phrase courte, renforce l'effet de contraste par rapport au phrases longues. Cet effet est d'ailleurs encore plus prononcé dans le deuxième exemple. Ici, de manière intéressante, un point final sépare la première phrase principale de sa subordonnée, introduite par *tant et si bien que*. La ponctuation proposée par l'auteur·e ne correspond donc pas à la structure syntaxique de la phrase, créant un effet de contraste et mettant en relief la subordonnée plutôt courte (17 mots). Ce procédé, stylistiquement très marqué, n'a pu être trouvé que dans le corpus portant sur les études littéraires.

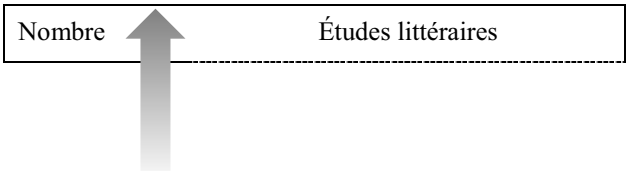
> Le ROTEM® comme le TEG® offrent dans ce cadre une évaluation rapide et globale des troubles de l'hémostase, permettant de définir des stratégies de traitement ciblé en utilisant des concentrés de facteurs afin de réduire l'utilisation des PSL et peut-être les effets secondaires qui les accompagne. Néanmoins, des travaux prospectifs devront confirmer ces hypothèses. La principale indication des techniques viscoélastique (TVE) est de diagnostiquer rapidement les troubles de l'hémostase et de guider leur traitement dans des contextes où une coagulopathie est fortement probable et/ou le risque d'hémorragie important (corpus biochimie_9).

> L'acteur remplacé par une installation plastique, la voix par un enregistrement sonore : ces inclinations montrent bien, dans la poétique novarinienne, à quel point l'incarnation reste une question délicate. Tant et si bien que parfois l'acteur laisse tout bonnement place à une étrange marionnette électronique. Cette marionnette, fabriquée par Zaven Paré, plasticien féru de robotique et de nouvelles technologies, avait déjà été utilisée en 1999, mais avec le visage de Novarina, lors de la création du Théâtre des oreilles en anglais, à Los Angeles (*cf.* corpus littérature_10).

Ainsi, si la variation aléatoire interne aux articles reste proéminente, nos analyses des 39 phrases plus courtes que la valeur minimale de l'intervalle de fluctuation (*cf.* tableau 4) auront tout de même pu montrer que l'alternance entre phrases longues et courtes est relativement fréquente et qu'elle peut être en partie expliquée par – à tout le moins – trois effets stylistiques. Premièrement, les phrases courtes introduisent souvent un nouvel aspect, combiné dans certains cas à des éléments cataphoriques, comme certaines locutions adverbiales (p. ex. *comme suit*) ou encore, en particulier dans le cas des études littéraires, des questions rhétoriques annonçant les phrases à venir. Ces dernières, généralement plutôt longues, peuvent présenter des informations ou réflexions supplémentaires, confirmant à nouveau que les *processus d'introspection ou de réflexion* définis plus haut (*cf.* 4.1) favorisent bel et bien les phrases plutôt longues (*cf.* également Barr 2001 : 378). Deuxièmement, les phrases courtes marquent parfois la fin d'une réflexion. Troisièmement, l'alternance entre phases courtes et longues peut vouloir créer un effet de contraste. Dans le cas des études littéraires, nous avons même pu observer la séparation graphique d'une phrase principale de sa subordonnée, un procédé stylistiquement marqué. Soulignons finalement que les observations ponctuelles de ce dernier chapitre, qui soulignent les effets stylistiques possibles de l'alternance de la longueur des phrases, restent illustratives et ne prétendent à aucune représentativité statistique par rapport à la globalité des phrases courtes se trouvant dans les 80 articles. Ces résultats devront donc encore être validés sur la base d'un corpus plus étendu.

5. Conclusion

En conclusion, la présente étude aura permis une analyse systématique de la longueur des phrases en prenant pour la première fois en compte quatre domaines de spécialité, permettant des conclusions par rapport à la « division horizontale » (Kocourek 1991 : 34) du langage scientifique, qui jusqu'ici n'avait été que trop peu étudiée. Pour ce qui est du principal résultat de la présente étude, soulignons tout d'abord qu'elle aura effectivement pu montrer que le domaine de spécialité des auteur·e·s influence de manière significative la longueur des phrases mesurées dans notre corpus, ce qui confirme notre hypothèse initiale. Ce résultat corrobore dans le même temps l'observation générale de Kocourek (1991 : 73) quant à l'existence d'un style d'écriture propre à chaque domaine de spécialité, et il l'affine également pour les domaines spécifiques des sciences naturelles et des sciences humaines. Ce faisant, la hiérarchie illustrée dans la figure 5 est suggérée par nos données.



de mots par phrase	Sociolinguistique	
	Biochimie	Phonétique

Fig. 5. Hiérarchie du nombre de mots par phrase dans les quatre disciplines étudiées.

Ainsi, la différence de longueurs de phrases observée par Rinck (2006 : 189) dans les sciences du langage et littéraires est globalement confirmée. Nos résultats ajoutent cependant une nuance au sein des sciences du langage, cette différence pouvant, du moins en partie, expliquer la forte variation observée par Rinck (2006 : 190) au sein de cette discipline. En effet, les articles de phonétique présentent des phrases globalement plus courtes que ceux de sociolinguistique. Ceci confirme ainsi notre hypothèse initiale stipulant que les articles en phonétique pourraient avoir tendance à présenter des longueurs de phrases se rapprochant de celles des sciences naturelles, alors que celles de la sociolinguistique se localisent bien dans le cadre des sciences humaines. Soulignons également que cette observation se trouve par ailleurs en corrélation avec les différences de fréquence des points-virgules et des deux-points. Mais étant donné le faible nombre de phrases contenant ces deux signes de ponctuation (84/800), leur effet sur nos résultats reste faible. Par ailleurs, en plus de ces différences interdisciplinaires, nos résultats auront également pu montrer que la méthode de travail employée par les auteur·e·s peut également influer sur la longueur des phrases, ce qui peut ajouter des nuances à l'intérieur de ces disciplines. La description de faits empiriquement observés est en effet souvent faite sur la base de phrases plutôt courtes et accompagnée de nombreux éléments péritextuels, en général statistiques, alors que des processus d'introspection ou de réflexion semblent favoriser l'utilisation de phrases plutôt longues. Ainsi, à titre d'exemple, un·e biochimiste travaillant dans un domaine expérimental présentera plus fortement cette tendance aux phrases courtes qu'un·e biochimiste spécialisé·e dans les questions d'éthique. De manière générale, nos résultats confirment donc en partie le vieux stéréotype du style de rédaction plutôt bref et 'avare en mots' des sciences naturelles – lorsque ces dernières font référence à des données empiriquement observées – et du style plus étendu des sciences humaines, qui se basent plus souvent sur des processus d'introspection ou de réflexion.

Les résultats portant sur les effets aléatoires de notre modèle ont cependant également pu montrer une forte variation des longueurs de phrases, d'une part, entre les 80 articles (variance : 42,81 mots ; écart type : 6,54 mots) et, d'autre part, à l'intérieur même des articles (*cf.* fig. 4). Ce résultat confirme donc largement les résultats de Barr (2001 : 377) quant à la grande variation des longueurs de phrases dans les articles scientifiques. Nos analyses qualitatives complémentaires sur le cotexte des 39 phrases particulièrement courtes présentent par ailleurs de possibles explications stylistiques à cette variation. Nous avons en effet pu montrer – à titre d'exemple sur la base de ce petit nombre d'observations – trois effets stylistiques liés à l'utilisation régulière d'une alternance entre phrases plutôt courtes et phrases plutôt longues : les phrases courtes peuvent présenter une valeur introductive, conclusive ou contrastive. Par ailleurs, elles contiennent souvent de brefs constats suivis par une phrase plus longue contenant des explications ou des réflexions plus détaillées.

Le présent article aura également laissé plusieurs questions et perspectives de recherche ouvertes. Premièrement, il serait intéressant de découvrir dans quelle mesure l'influence de l'anglais, qui est souvent considéré comme une langue de publication favorisant de plus en plus les phrases plutôt courtes (*cf.* p. ex. Rudnicka 2018 : 233), est observable dans les publications d'auteur·e·s de disciplines dans lesquelles l'anglais est majoritairement reconnu comme étant la principale langue de publication (p. ex. en biochimie ou en phonétique ; *cf.* 3 : *Revues*). Deuxièmement, les résultats se devraient encore d'être vérifiés sur la base d'aspects non plus seulement graphiques mais syntaxiques, comme – entre autres – le type de phrase, la structure thème-rhème, la condensation syntaxique, l'impersonnalité, la nominalisation des prédicats et la désémantisation des verbes (*cf.* 3.2). Troisièmement, notre corpus de 800 phrases dans quatre disciplines reste relativement restreint, de sorte qu'une étude similaire se basant sur un plus large corpus et d'autres disciplines pourrait apporter des nuances supplémentaires à nos résultats. Finalement, de manière plus générale, il serait intéressant de découvrir dans quelle mesure les représentations de l'écriture scientifique idéalisé correspondent (ou non) aux productions concrètement observables dans les articles. Les études publiées jusqu'ici ont en effet généralement porté soit uniquement sur les représentations (*cf.* p. ex. Reutner 2008, 2009, 2013), soit uniquement sur les productions (*cf.* 2). Une comparaison directe de ces deux niveaux pourrait donc s'avérer fructueuse.

Références bibliographiques

Barr, G. (2001). Graphical Analysis of the Sentence Length Distribution Curve and Non-rational Components. *Literary and Linguistic Computing*, 16/4, 375–388.

Bennett, K., Muresan, L. (2016). Rhetorical incompatibilities in academic writing : English versus the Romance Languages. *Synergy*,12/1, 95–119.

Best, K.-H. (2005). Satzlänge, dans : Köhler, R., Altmann, G./Piotrowski, R. G. (éds.) : *Quantitative Linguistik*. *Ein internationales Handbuch*. Berlin/New York : de Gruyter, 298–304.

Clyne, M. (1991). The Sociocultural Dimension : The Dilemma of the German-speaking Scholar, dans : Schröder, H. (éd.) : *Subject-oriented texts : language for special purposes and text theory*. Berlin/New York : de Gruyter, 49–67.

Kelih, E./Grzybek, P. (2004). Satzlänge: Definitionen, Häufigkeiten, Modellierung (am Beispiel slowenischer Prosatexte). *LDV-Forum. Zeitschrift für Computerlinguistik und Sprachtechnologie* 20, 31–51.
Gray, B./Biber, D. (2018). Academic writing as a locus of grammatical change. The development of phrasal complexity features, dans : Whitt, R. J. (éd.). *Diachronic Corpora, Genre, and Language Change*. Amsterdam/Philadelphia : John Benjamins, 117–146.
Fengxiang, F. (2007). A corpus based quantitative study on the change of TTR, word length and sentence length of the English language, dans : Grzybek, P., Köhler, R. (éds.). *Exact Methods in the Study of Language and Text*. *Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*. Berlin/New York : Mouton De Gruyter, 123–130.
Fifielska, E. (2015). Les constructions syntaxiques de l'écrit scientifique: exploration et analyses de corpus, disponible sur https://dumas.ccsd.cnrs.fr/dumas-01213405 (10.12.2019).
Galtung, J. (1983). Struktur, Kultur und intellektueller Stil. Ein vergleichender Essay über sachsonische, teutonische, gallische und nipponische Wissenschaft. *Leviathan* 11/3, 303–338.
Genette, G. (1987). *Seuils*. Paris : Éditions du Seuil.
Hoffmann, L. (1998). Syntaktische und morphologische Eigenschaften von Fachsprachen, dans : Steger, H., Wiegand, H. E. (éds.). *Fachsprachen*. Berlin/New York : de Gruyter, 416–427.
Karlgren, J., Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis, dans : Nagao, M. (éd.) : *Proceedings of COL-ING 94*. Kyoto, 1071–1075.
Kelih, E., Grzybek, P., Antić, G., Stadlober, E. (2006). Quantitative Text Typology. The Impact of Sentence Length, dans : Spiliopoulou, M./Kruse, R./Borgelt, C./Nürnberger, A./Gaul, W. (éds.) : *From Data and Information Analysis to Knowledge Engineering. Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Magdeburg, March 9-11*, 2005. Berlin : Springer, 382–389.
Kocourek, R. (^{2}1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. Wiesbaden : Oscar Brandstetter.
Larivière, V. (2019). Le français, langue seconde ? De l'évolution des lieux et langues de publication des chercheurs québécois, français, et allemands, https://crctcs.openum.ca/files/sites/60/2019/08/Langues_Rech_Soc_Revise.pdf (13.12.2019).
Monjallon, A. (1980). Introduction à la méthode statistique. Paris : Vuibert.
Niemann, R. (2018). *Wissenschaftssprache praxistheoretisch. Handlungstheoretische Überlegungen zu wissenschaftlicher Textproduktion*. Berlin/Boston : de Gruyter.
Paillé, P., Mucchielli, A. (2012). *L'analyse qualitative en sciences humaines et sociales*. Paris : Armand Colin.
Petkova-Kessanlis, M. (2015). Nachfeldbesetzungen und ihre kommunikative Funktion in wissenschaftlichen Texten, dans : Vinckel-Roisin, H. (éd.). *Das Nachfeld im Deutschen : Theorie und Empirie*. Berlin/Boston : de Gruyter, 211–228.
Pontille, D. (2003). Formats d'écriture et mondes scientifiques, *Questions de communication*, 3, 55–67.
Reutner, U. (2008). Le 'bon usage' scientifique. Une enquête menée dans le domaine de la linguistique, dans : Reutner, U./Schwarze, S. (éds.) : *Le style, c'est l'homme. Unité et diversité du discours scientifique dans les langues romanes*. Frankfurt-sur-le-Main et al. : Peter Lang, 249–284.
Reutner, U. (2009). Aspetti sintattici del discorso scientifico: risultati di un'inchiesta, dans : Ferrari, A. (éd.): *Sintassi storica e sincronica dell'italiano. Subordinazione, coordinazione, giustapposizione. Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana (Basilea, 30 giugno - 3 luglio)*. Florence : Cesati, 1409–1428.
Reutner, U. (2013). La tridimensionalidad de la transmisión del saber : culturas nacionales, disciplinarias y graduales, dans : Sinner, C. : *Comunicación y transmisión del saber entre lenguas y culturas.* München: Peniope (Études linguistiques/Linguistische Studien 10), 443–463.
Rinck, F. (2006). *L'article de recherche en Sciences du langage et en Lettres*. *Figure de l'auteur et identité disciplinaire du genre*, disponible sur http://ed.humanites.unistra.fr/uploads/media/thesefannyrinck.pdf (11.12.2019).
Rouleau, M. (2006). Longueur comparée de la phrase médicale et de la phrase générale, *Équivalences*, 33/1–2, 137–147.
Rudnicka, K. (2018). Variation of sentence length across time and genre: influence on the syntactic usage in English, dans : Whitt, Richard Jason (éd.) : *Diachronic Corpora, Genre, and Language Change.* Amsterdam/Philadelphia : John Benjamins, 220–240.
Sigurd, B./Eeg-Olofsson, M./van de Weijer, J. (2004). Word Length, Sentence Length and Frequency – Zipf revisited, *Studia Linguistica*, 58/1, 37–52.
Simmler, F. (2006). Varietätenlinguistik : Fachsprachen, dans : Ágel, V. *et al.* (éds.). *Dependenz und Valenz. Ein internationales Handbuch der zeitgenössischen Forschung*. Berlin/New York: de Gruyter, 1523–1538.
Snow, C.P. ([1959] 2018). *The two cultures and the scientific revolution*. Cambridge : Cambridge University Press.
Torttila, M., Hakkarainen, H. J. (1990). Zum Satzbau der deutschen Kochrezepte des 20. Jahrhunderts : Satzlänge und Prädikat, dans : *Zeitschrift für Germanistische Linguistik* 18/1, 31–42.
Trubetzkoy, N.S. ([1939] ^{7}1989) : *Grundzüge der Phonologie*. Göttingen : Vandenhoeck & Ruprecht.

Extraits du corpus

Corpus biochimie_3 = Levine, E., Beroul, N., Cortey, A., Damais Cepitelli, A., Gouezec, H., Pujol, S., Wibaut, B., Marti. B. (2018). Réalisation d'une transfusion sanguine en néonatologie : état des lieux des pratiques en 2016 en France ; situations aiguës ou chirurgicales exclues. *Transfusion Clinique et Biologique* 25, 249–256.
Corpus biochimie_9 = David, J.S., Imhoff, E., Parat, S., Augey, L., Geay-Baillat, M.-O., Incagnoli, P., Tazarourte, K. (2016). Intérêt de la thromboélastographie pour guider la correction de la coagulopathie post-traumatique : plus de MDS, moins de PSL? *Transfusion Clinique et Biologique* 23, 205–211.
Corpus littérature_1 = Denizot, M. (2014). Robespierre de Romain Rolland : Une écriture de l'histoire par analogie. *Études théâtrales*, 59/1, 73–84.
Corpus littérature_2 = De Simone, C. (2014). Pinocchio d'après Carmelo bene : Langue de bois et voix de chair pour la réinvention d'une marionnette. *Études théâtrales* 60–61, 63–72.
Corpus littérature_6 = Dubor, F. (2014). Portrait du théâtre, corps et voix, en marionnette. *Études théâtrales* 60–61/2, 102–111.
Corpus littérature_10 = Hersant, C. (2014). Médiatisation du corps et des discours : l'acteur novarinien et ses doubles. *Études théâtrales* 60–61/2, 154–162.

Corpus_littérature_11 = Thiériot, G. (2014). Les premières pièces de Werner Schwab. Une autre façon de tirer les ficelles ? *Études* théâtrales 60–61, 147–153.
Corpus phonétique_9 = Broux, P.-A., Doukhan, D., Petitrenaud, S., Meignier S., Carrive J. (2018). Segmentation et Regroupement en Locuteurs: comment évaluer les corrections humaines, dans : Cooke, M., Bigi, B., Lavaud, J. (éds.). *Actes de la XXXIIe Journées d'Études sur la Parole 4-8 juin 2018*, Aix-en-Provence, France, 89–97.
Corpus phonétique_11 = Cattelain, M., Savariaux, C., Gerber, S., Perrier, P. (2018). Analyse électromyographique de la production des plosives labiales : enjeux méthodologiques, dans : Cooke, M., Bigi, B., Lavaud, J. (éds.). *Actes de la XXXIIe Journées d'Études sur la Parole 4-8 juin 2018*, Aix-en-Provence, France, 107–115.
Corpus phonétique_13 = Yamilamai, N., Tran, T. T. H. (2018). Effet de la position de la syllabe sur la réalisation acoustique des consonnes finales du thaï, dans : Cooke, M., Bigi, B., Lavaud, J. (éds.). *Actes de la XXXIIe Journées d'Études sur la Parole 4-8 juin 2018*, Aix-en-Provence, France, 151–159.
Corpus phonétique_15 = Mdhaffar, S., Laurent, A., Estève, Y. (2018). Etude de performance des réseaux neuronaux récurrents dans le cadre de la campagne d'évaluation Multi-Genre Broadcast challenge 3 (MGB3), dans : Cooke, M., Bigi, B., Lavaud, J. (éds.). *Actes de la XXXIIe Journées d'Études sur la Parole 4-8 juin 2018*, Aix-en-Provence, France, 169–177.
Corpus sociolinguistique_10 = Calvet, L.-J. (2016), Pratiques des langues en France. Oui, mais de quoi parlons-nous ? *Langage et société* 155, 39–59.

[1] Cette définition graphique prédomine dans la linguistique quantitative, principalement parce qu'elle fournit des critères d'analyse clairs (*cf*. Kelih/Gryzbek 2004 : 33). Elle n'est cependant pas sans poser des problèmes parce qu'elle se base sur des signes de ponctuation, alors que dans certains cas, comme par exemple dans certains poèmes du XX^{e} siècle, les frontières de la phrase ne peuvent être délimitées par de tels signes. Ainsi, les résultats basés sur ces critères pourront dans certains cas contenir plutôt des informations sur les pratiques de ponctuation que sur le nombre de phrases ou sur leur longueur.
[2] Conformément à Genette (1987, 10*sq.*), nous entendons par *péritexte* une sous-catégorie du *paratexte*, qui regroupe les éléments entourant et prolongeant le texte. Dans le *péritexte*, ces éléments sont situés à l'intérieur du texte, au contraire de l'*épitexte*, dans lequel ils se trouvent à l'extérieur du texte. Ainsi, le *péritexte* comprend des éléments comme les titres, sous-titres, intertitres, les noms des auteurs et éditeurs, la date d'édition, le préface, les notes (de bas de page), les tableaux, les graphiques, les images et illustrations ou encore la table des matières. L'*épitexte*, pour sa part, regroupe des catégories comme les interviews, les comptes-rendus journalistiques ou encore les correspondances de l'auteur.
[3] Nous définissons ici les phrases introduisant des éléments péritextuels comme étant des phrases mentionnant explicitement cet élément pour la première fois dans l'article, soit en mentionnant le tableau ou le graphique directement dans la phrase et en le commentant, soit en y faisant référence dans une parenthèse placée en fin de phrase.
[4] Par *éléments péritextuels statistiques*, nous entendons des tableaux ou graphiques contenant des données quantitatives – c'est-à-dire des données pouvant être mesurées ou repérées (*cf*. Monjallon 1980) – ainsi que des interprétations ou visualisations (également quantitatives) de ces données.
[5] Notons que l'unique illustration se trouvant dans le sous-corpus des études littéraires n'est pas introduite explicitement dans le texte, raison pour laquelle aucun chiffre ne peut être fourni ici.