La loi de Menzerath appliquée à un ensemble de textes

Marc Hug
Professeur émérite, Université Marc Bloch, 23, rue Descartes, 67084 Strasbourg Cedex

ABSTRACT : The so-called "Menzerath law" is well known by German linguists dealing with quantitative linguistics, while it is almost ignored by French linguists. It says that in any linguistic construct, the more constituents a given unit is made of, the smaller they are. Here this "law" is confronted with observations made on a corpus of texts taken from the newspaper Le Monde. It appears that globally articles (texts) that have long sentences use longer words than articles having shorter sentences - but inside most of these articles, long sentences use shorter words than short sentences. Similar measures made with an intermediate level between "sentence" and "word", or made on phonetic transcriptions of the texts, do not change the conclusions. This shows that a research on the Menzerath law should always provide precise information on the exact nature of the units that are being counted and the way they have been obtained.

KEYWORDS : Menzerath law, sentence, clause, word, punctuation, linear regression, linear correlation.

RESUME : La loi dite de Menzerath est bien connue des linguistes allemands faisant de la linguistique quantitative, mais presque totalement ignorée de leurs collègues français. Elle dit que "plus une unité linguistique d'un niveau donné comporte de constituants du niveau inférieur, plus ces constituants sont petits". Cette "loi" est ici confrontée aux observations faites sur un ensemble de textes tirés du quotidien Le Monde. On constate que globalement les articles qui ont les phrases les plus longues font aussi usage des mots en moyenne plus longs que les articles dont les phrases sont plus courtes - mais qu'à l'inverse, à l'intérieur de la plupart de ces articles, les phrases les plus longues ont des mots en moyenne plus courts que les phrases courtes. Des mesures analogues faites en prenant en compte un niveau intermédiaire entre la phrase et le mot, ou en opérant sur les transcriptions phonétiques des textes, ne conduisent pas à des conclusions différentes. Il apparait que si l'on travaille sur la loi de Menzerath, il est indispensable de dire très précisément quelles unités on a comptées et de quelle façon précise on a obte nu les données sur lesquelles on calcule.

MOTS CLES : Loi de Menzerath, phrase, proposition, syntagme, mot, ponctuation, régression linéaire, corrélation linéaire.

La loi de Menzerath

Les linguistes français, même ceux qui sont un peu frottés de linguistique quantitative, ne connaissent guère la loi dite de Menzerath, ou de Menzerath-Altmann, qui est au contraire fort connue parmi les linguistes allemands et ceux des linguistes anglophones qui sont au contact de la linguistique allemande. Paul Menzerath (1883-1954)[1] fut un phonéticien allemand qui publia en 1954 un ouvrage de statistique lexicale relatif aux structures statistiques du lexique allemand, *Die Architektonik des deutschen Wortschatzes*, où est exprimée la "loi" qui porte son nom. Gabriel Altmann est actuellement la personnalité la plus en vue dans la linguistique quantitative en Allemagne. La "loi de Menzerath" (Menzerath 1954:101), baptisée ainsi, popularisée et étudiée par Gabriel Altmann, dit que "plus le tout est grand, plus ses constituants sont petits". Cette loi s'entend pour tout ensemble linguistique constitué d'un nombre variable de constituants d'un rang inférieur. Dans le *mot*, le constituant plus petit sera la syllabe (ou le phonème, ou le morphème), et la loi se traduira par la règle particulière (a) "plus le mot comporte de syllabes, moins les syllabes comporteront, en moyenne, de consonnes" (sachant qu'en principe une syllabe se définit par la présence d'une voyelle et une seule). Si l'ensemble superordonné est le *syntagme*, la *proposition* ou la *phrase*, la partie pourra être le mot, et la règle particulière (b) sera de la forme : "plus la phrase comporte de mots, plus la longueur moyenne de ces mots sera réduite". On voit d'emblée que si la règle particulière (a) ne soulève pas trop d'objections, la règle (b) parait au contraire beaucoup plus sujette à contestation. Rien ne dit en effet que la "loi de

1. Il n'est par certain que Paul Menzerath ait vu la parution de son livre de 1954. Il est mort le 8 avril 1954. Il était né le 1er janvier 1883 à Düren (à l'Est d'Aix-la-Chapelle). Merci au Dr. Thomas Mauersberg, de l'Université de Bonn, pour ces précisions.

Menzerath" s'applique directement aux mots en tant que constituants de la phrase. G. Altmann (1989:;11-12), dans un ouvrage collectif, dit même expressément que pour lui, le syntagme est un constituant de la phrase, le mot un constituant du syntagme, et que par conséquent, plus la phrase est longue par le nombre de ses syntagmes, plus les syntagmes doivent être courts par le nombre des mots, et par conséquent plus, dans ces syntagmes plus courts, les mots doivent être plus longs. Il est vrai qu'il prend la précaution de prévoir ici bien des facteurs perturbants (*Störfaktoren*). De toute façon, la description implicite qu'il fait de la phrase dans ce contexte est des plus sommaires, et on peut imaginer - comme on le fait par ailleurs dans le même ouvrage - que la "proposition" (*Teilsatz*) est un autre constituant de la phrase, composé lui-même de syntagmes, ce qui introduit un niveau hiérarchique intermédiaire. Mais ces "subordonnées" peuvent-elles être découpées selon la méthode bouchère à l'ancienne (*Je crois / que tu devrais acheter les lunettes / que l'oculiste t'a prescrites*) ? Ou bien la "principale" inclut-elle les subordonnées ? Les syntagmes constituent-ils un seul niveau, ou bien faut-il distinguer des niveaux de syntagmes dans chaque "proposition" ? En l'absence de recherches fondées sur une analyse linguistique explicite, il est difficile de se prononcer sur les niveaux de constituants dont la distinction manifesterait le plus clairement l'influence de la loi de Menzerath.
La question qu'on voudrait soulever ici est celle des conditions d'observation et de mesure qu'on se donne pour étudier cette loi. Le point de départ de cette réflexion a été fourni par les remarques critiques faites par Maria Roukk (2003). L'application particulière à propos de laquelle Paul Menzerath a eu l'intuition d'une telle règle est une étude sur le *lexique* de l'allemand, à propos du nombre de syllabes des mots et du nombre de phonèmes des syllabes. Il examine donc l'inventaire des unités lexicales disponibles - en fait un sous-ensemble d'unités présentes dans un dictionnaire, à quoi il ajoute selon une méthode relativement peu rigoureuse un certain nombre de lexèmes supplémentaires. Toujours est-il qu'il s'agit d'un inventaire d'unités, et que Menzerath dit et répète que la fréquence d'emploi des unités ne fait pas partie, pour lui, de la "structure" de la langue, et qu'elle ne l'intéresse pas, moyennant quoi il peut arriver à la conclusion que le mot allemand le plus "fréquent" est le mot de trois syllabes : il veut dire qu'il existe plus de mots de cette longueur que de mots d'autres longueurs. Or les études qui ont été consacrées à sa loi concernent souvent des textes, dans lesquels les unités répétées sont comptées à chacune de leurs occurrences. Il ne s'agit donc plus du tout de la même loi, malgré l'avis contraire de G. Altmann (cf. ci-dessous). Il est facile de comprendre pourquoi on a dérivé de cette façon : d'une part plus personne, en tout cas parmi les quantitativistes, ne dirait que la fréquence d'emploi des vocables ne fasse pas partie des structures linguistiques. D'autre part l'attraction des études de G. K. Zipf, qui portait sur des fréquences d'emploi, a fait qu'on a interprété aussi la loi de Menzerath en termes de paramètres applicables aux textes.
Cela posé, le problème se présente dans les termes suivants : des études assez nombreuses déjà ont montré que dans la comparaison entre les langues, la loi de Menzerath s'applique bien en ce sens, par exemple, qu'une langue comme le latin, usant d'un nombre relativement restreint de mots-outils grammaticaux, a des mots en moyenne plus longs et des phrases composées de moins de mots qu'une langue de type beaucoup plus analytique telle que l'anglais. Mais d'un autre côté, nous pouvons avoir l'impression qu'à l'intérieur d'une même langue, deux tendances antagonistes pourraient bien se faire jour :
- d'une part il peut y avoir dans la langue des tendances plus ou moins analytiques selon le "style", et la loi de Menzerath pourrait donc s'appliquer dans la comparaison entre textes ;
- d'autre part il pourrait y avoir, selon la nature des référents, une nécessité d'avoir dans certains cas à la fois des mots brefs et des phrases brèves, dans d'autres des unités plus longues, ce qui pourrait produire un effet contraire à celui que prévoit cette loi.

Le corpus étudié

Au lieu de spéculer sur des intuitions, voyons plutôt ce qu'on peut constater sur une série particulière de textes. J'ai travaillé ici sur un ensemble d'une centaine d'articles du quotidien *Le Monde*, qui présente cet avantage inestimable de fournir du texte récent sur un support informatique. Les textes extraits ne constituent pas à proprement parler un "échantillon représentatif" au sens technique de l'expression, mais on peut les considérer comme raisonnablement représentatifs tout de même : il s'agit de l'ensemble constitué par trois séries extraites successivement :
- une première série de 22 articles du numéro daté du 4 juin 1999 ;
- une deuxième série de 36 articles extraits des premières pages des numéros des 18, 21, 23, 27 et 30 mars 2002 ;
- enfin une troisième série de 45 articles extraits des pages 7 et suivantes du numéro du 28 mars 2002 et des six premières de celui du 29 mars 2002.

Dans tous les cas, on a pris soin d'éliminer les articles trop courts - sur un critère intuitif : on éliminait les articles qui s'affichaient en totalité sur un même écran d'ordinateur, ce qui nous faisait écarter la plupart des articles de moins de 600 mots, ou 3300 caractères. Au bout du compte on a un ensemble d'articles qui, du point de vue de la longueur, ne sont pas répartis selon une courbe vraiment gaussienne, faute de symétrie : la

longueur moyenne est de l'ordre de 900 mots, mais les articles les plus longs dépassent les 2000 mots, alors que le plus court en comporte 406. La tranche la mieux représentée est celle des articles de 600 à 1000 mots, qui constituent plus de la moitié à eux seuls (58 sur 103).

Définition de la phrase et du mot

Les deux niveaux sur lesquels j'ai travaillé sont deux niveaux qu'on peut considérer comme éloignés l'un de l'autre : celui du mot et celui de la phrase. J'ai étudié le rapport qui peut exister entre le nombre de mots dont se compose la phrase et le nombre de lettres écrites, resp. de phonèmes, dont se compose le mot.
Je partais d'une définition du mot telle qu'elle est pratiquée dans la base FRANTEXT : un mot est toute suite continue de caractères alphabétiques ; l'apostrophe est considérée comme un signe alphabétique en ce sens qu'il entre dans la composition du mot qui le précède, mais elle est toujours considérée comme une limite de mot (dernière lettre du mot) ; tout trait d'union, tout signe de ponctuation, parenthèse etc., tout chiffre est considéré comme un mot. Les points de suspension sont considérés comme un seul mot. Par extension, j'ai considéré aussi comme un seul mot les cas exceptionnels dans lesquels plusieurs points d'interrogation ou d'exclamation se suivent.
La limite de phrase a été fixée de manière aussi simpliste : tout point, point d'interrogation, point d'exclamation, point-virgule ou deux-points est considéré comme une fin de phrase, et la ponctuation en question est le dernier mot de la phrase.
Il va de soi que ces définitions sont linguistiquement peu satisfaisantes. Mais en s'inspirant de la règle qui veut que "l'essentiel est que cette norme soit constante" (Muller 1973: 10), on peut se dire que les conclusions seront tout de même valables, puisque le traitement informatique nous garantit la constance de la norme. Il sera tout de même prudent de s'en assurer.
Tout de même, on a cru utile d'intervenir sur les textes pour éviter de trouver des "fins de phrase" chaque fois qu'un prénom est noté par son initiale suivie d'un point, ou lorsqu'une personnalité est désignée par "*M. ...*" (*Monsieur*) Le point qui suit *etc.* est en revanche peu gênant, car la plupart du temps cette abréviation se rencontre en fin de phrase. On n'a pas touché non plus aux mots *aujourd'hui* et *quelqu'un*, qui ont donc été coupés en deux "mots".

Moyennes

Dans un premier temps, un programme informatique a calculé dans chacun des 103 textes sélectionnés
- le nombre moyen de mots par phrase
- le nombre moyen de lettres par mot
en appliquant les règles qui viennent d'être énoncées.
Mais pour s'assurer que les particularités de la norme ne biaisent pas trop sensiblement les résultats, ce calcul a été fait de deux façons :
- d'une part en comptant effectivement tous les mots comme il a été dit, y compris les signes de ponctuation ;
- d'autre part en laissant de côté les ponctuations.
En effet, comme ces signes marquent par définition la fin de chaque phrase, ils constitueront d'emblée un mot de 1 lettre (en général) dans chacune des phrases, quelle que soit par ailleurs sa longueur. De ce fait, la longueur moyenne du mot sera fortement diminuée dans le cas d'une phrase très courte (de très peu de mots), beaucoup moins dans le cas d'une phrase longue. Ceci est, de toute évidence, de nature à fausser le résultat d'un calcul relatif à la loi de Menzerath. Prenons le cas imaginaire d'une phrase qui serait composée, par ailleurs, uniquement de mots de deux lettres et de mots de six lettres en nombre égal. Dans une phrase de 2 mots suivis d'un point (Ex. *Il arrive.*), j'obtiendrai, en comptant le point, un total de 9 caractères faisant trois mots, donc une longueur moyenne du mot de 3 lettres. Si je ne compte pas le point, la longueur moyenne sera de 4 lettres par mot, évidemment. A supposer maintenant que j'aie une phrase avec 7 mots de deux lettres et 7 mots de six lettres, suivis d'un point, j'aurai au total, en comptant le point, 15 mots totalisant 57 caractères, donc une moyenne de 3,8 lettres par mot, alors que si je ne compte pas le point, j'aurai toujours une moyenne de 4. Dans la comparaison de ces deux phrases, j'aurai peut-être l'impression, si j'ai compté le point, que la phrase plus longue a des mots plus longs que la phrase plus courte, ce qui est contraire à la loi de Menzerath ; si je n'ai pas compté le point, je ne trouverai aucune incidence de la longueur de la phrase sur la longueur du mot.
Il est entendu que les comparaisons qu'on peut faire entre les deux moyennes obtenues sur les différents textes sont un peu biaisées ici par le fait que les textes ne sont pas tous de même longueur, et que par conséquent, si on suppose tous les textes également homogènes, les moyennes présentent dans les textes les plus courts une variabilité aléatoire plus grande que dans les textes les plus longs. Mais vu le nombre de textes examinés, il est permis de penser que les effets de ce biais sont à peu près insensibles au total. Ce qui est plus important sans doute, on le verra plus loin, c'est que les deux moyennes ainsi calculées seront en fait des paramètres *du texte* et non des paramètres des phrases ou des mots.

Corrélations entre textes

A partir des deux moyennes (longueur du mot, longueur de la phrase) calculées à propos de chaque texte, on a calculé sur l'ensemble des 103 couples de données un coefficient de corrélation linéaire.
On obtient les résultats suivants :
- si l'on ne compte pas les signes de ponctuation,

$r = 0{,}24658$, ou avec l'approximation normale $u = 2{,}518$.

$R = 0{,}21444$, ou avec l'approximation normale $u' = 2{,}178$.

(Je note r le coefficient de corrélation linéaire, R le coefficient de corrélation des rangs).
On obtient donc une corrélation positive significative : les articles ayant des phrases relativement longues tendent à être constitués aussi de mots plus longs. Si le mot peut être considéré comme constituant de la phrase, on fait donc une observation opposée à ce que ferait prévoir la loi de Menzerath. L'explication qu'on peut proposer - mais qu'il sera difficile de fonder sur des faits indiscutables - est que certains articles sont sur des sujets moins techniques que d'autres, et demandent à la fois des phrases moins longues et un vocabulaire moins spécialisé. Les mots les plus courants, on le sait, sont parmi les mots courts.
Si nous comptons les signes de ponctuation, les résultats correspondants seront les suivants :

$r = 0{,}39556$, ou avec l'approximation normale $u = 4{,}184$.

$R = 0{,}35782$, ou avec l'approximation normale $u' = 3{,}744$.

Comme notre exemple imaginaire nous ne faisait prévoir, la tendance opposée à la loi de Menzerath se manifeste encore plus nettement ici, mais on peut craindre qu'il ne s'agisse d'un artefact, même si les virgules présentes plus nombreuses dans les phrases plus longues atténuent forcément le biais introduit par la prise en compte du point final.

Corrélation interne aux textes

Mais il y a moyen de calculer d'autres coefficients de corrélation : à l'intérieur de chaque texte, nous avons des phrases, et chaque phrase est composée de mots. Nous pouvons dresser une liste de mesures relatives à chaque phrase :
- le nombre de mots de la phrase ;
- le nombre moyen de lettres dont se composent les mots de la phrase (nombre total des lettres divisé par le nombre des mots).

A partir de cette liste de couples de mesures, nous pouvons calculer pour le texte un coefficient de corrélation. Si nous faisons ce calcul pour l'ensemble des 103 textes, nous aurons 103 coefficients de corrélation.
Sur ces 103 coefficients, si nous ne comptons pas les ponctuations, nous en avons 74 qui sont négatives, les 29 autres qui sont positives. Si nous comptons les ponctuations, nous en avons 50 qui sont négatives et les 53 autres positives.
Nous pouvons considérer sans hésitation que le résultat obtenu sans les ponctuations est linguistiquement plus pertinent que l'autre. C'est aussi celui qui manifeste la contradiction la plus nette par rapport au résultat précédent. Dans les derniers calculs, la loi de Menzerath semble se vérifier, puisque le plus souvent, à l'intérieur d'un article donné, les phrases les plus courtes ont des mots plus longs en moyenne que les phrases plus longues.
Comment peut-on interpréter linguistiquement ce fait, et le concilier avec ce qu'on a observé précédemment ? Toujours sans possibilité de preuve factuelle indiscutable, on peut avancer que si les différents textes peuvent s'opposer - même à l'intérieur d'un genre particulièrement cohérent comme la prose journalistique du *Monde* - par des degrés de technicité qui influent sans doute parallèlement sur la longueur moyenne du mot et sur celle de la phrase, au contraire à l'intérieur d'un article donné, qui garde tout au long son niveau de technicité, forte ou faible, une loi linguistique plus profonde peut se manifester sans être autant contrariée par des oppositions qu'on peut qualifier de "stylistiques".

Prise en compte d'un niveau de "syntagmes"

Dans les travaux déjà réalisés, on ne compte pas directement, en général, le nombre des mots par phrase, et G. Altmann, s'appuyant sur Arens (1965), suggère que la tendance pourrait être inversée lorsqu'on saute un niveau hiérarchique de constituants. Cette question est difficile, et une des questions à résoudre serait de savoir quels sont précisément les niveaux hiérarchiques à distinguer.
Mais nous avons voulu vérifier si la prise en compte d'un niveau intermédiaire change quelque chose à nos conclusions. Pour ce faire, nous avons découpé les phrases, non selon les "propositions" de l'analyse grammaticale d'autrefois, ni selon les niveaux de l'analyse syntagmatique, mais simplement d'après la ponctuation : à chaque virgule ou tiret (distingué du trait d'union), on rencontre une limite de groupe. Mis à part que c'est le découpage le plus aisé à faire par des moyens informatiques, il faut ajouter que c'est un

critère plutôt prosodique que syntagmatique, et dans la mesure où la règle de Menzerath est liée à la prosodie, un tel critère est le bienvenu. Pour ne pas avoir l'air de vouloir induire en erreur, nous parlerons de "groupes" et non de syntagmes.
Quelle corrélation observe-t-on dans les moyennes globales des 103 textes entre le nombre de groupes par phrase, de mots par groupe et de lettres par mot ? les coefficients de corrélation sont les suivants (avec, toujours, 101 degrés de liberté) :
- groupes par phrase / mots par groupe : $r = 0{,}104$ - Approximation normale $u = 1{,}044$
- mots par groupe / lettres par mot : $r = 0{,}190$ - Approximation normale $u = 1{,}928$
- groupes par phrase / lettres par mot : $r = 0{,}251$ - Approximation normale $u = 2{,}567$.

(Le test de corrélation entre le nombre de mots par phrase et le nombre de lettres par mot n'a pas été replacé ici)
Comme précédemment, la corrélation est ici positive, sans atteindre toutefois le seuil usuel de 5 % dans la corrélation entre deux niveau voisins.
Faisons maintenant, comme précédemment, le test des corrélations observées à l'intérieur de chacun des textes ; nous aurons trois niveaux de données par texte :
(a) Nombre de groupes par phrase (un effectif par phrase)
(b) Nombre de mots par groupe (une moyenne par phrase, ou un effectif dans le cas de la phrase à un seul groupe)
(c) Nombre de lettres par mot (une moyenne par phrase ; un effectif seulement lorsque la phrase se compose d'un seul mot).
Pour chacune des variables (a), (b) et (c), nous avons autant d'expressions que de phrases dans le texte, et nous pourrons calculer la corrélation entre deux des variables dans chacun des textes. Le même calcul étant répété pour chacun des 103 textes, nous obtenons trois séries de 103 coefficients de corrélation.
Voici la répartition qu'on obtient entre corrélations négatives et corrélations positives :
- Série I, groupes par phrase / mots par groupe (a et b): 102 négatives, 1 positive
- Série II, mots par groupe / lettres par mot (b et c) : 70 négatives, 33 positives
- Série III, groupes par phrase / lettres par mot (a et c) : 52 négatives, 51 positives.

Alors que dans les deux premières séries, la disproportion est significative, l'équilibre est au contraire parfait dans la troisième. Du moment que le troisième ensemble de corrélations a été obtenu sur des données assez éloignées les unes des autres, il est normal que le lien ne se fasse pas sentir ; on remarque tout de même que si la "transitivité" envisagée par Altmann (cf. ci-après) existe, il faut penser qu'ici elle est complètement masquée par des facteurs perturbants. L'aspect le plus intéressant de cet ensemble de résultats est la corrélation à peu près toujours négative entre le nombre de groupes par phrase et le nombre de mots par groupe. Ce résultat justifie a posteriori le choix du critère de délimitation retenu.
Il faut toutefois ajouter aussitôt à cette appréciation un *bémol* : il peut y avoir des auteurs différents qui, dans le même cas, mettent ou ne mettent pas la virgule. On trouve aujourd'hui des personnes qui mettent systématiquement une virgule après la conjonction *or*, d'autres qui n'en mettent que rarement entre une principale et une subordonnée circonstancielle. Si, pour une même phrase, un auteur met trois virgules, un autre cinq, il est clair que la corrélation entre le nombre des "groupes" et leur longueur sera négative. Il ne me semble pourtant pas que ces habitudes divergentes jouent un rôle déterminant dans le résultat qu'on vient d'obtenir. Pour en avoir le cœur net, il faudrait normaliser la ponctuation des textes utilisés, ou compléter ce travail par la prise en compte d'autres types de constituants, tels que les "propositions" au sens de l'analyse logique ancienne, ou un autre niveau à définir. Si l'on veut utiliser les "constituants immédiats" de l'analyse syntagmatique, il sera sans doute nécessaire de comparer le nombre de niveaux syntagmatiques et la longueur des segments terminaux, qui sont en principe les mots.

Traitement de textes oraux

On sait que la norme écrite du français est fort éloignée de la forme orale de la langue, et si l'on pense que, par exemple, le mot *fini* se compose de quatre lettres à l'écrit et de quatre phonèmes à l'oral, cependant que le mot *temps* comporte cinq lettres à l'écrit, et presque toujours deux phonèmes seulement à l'oral, il y a lieu de douter de la représentativité d'un corpus écrit pour traduire les lois applicables à la langue orale. Pour un simple essai, j'ai fait la transcription phonétique de dix de mes textes (numéros choisis au hasard parmi les 103 précédemment traités sous forme écrite), au moyen d'un programme informatique qui propose pour chaque mot déjà rencontré la transcription retenue précédemment, en donnant cependant la possibilité d'en changer.
Bien entendu, le procédé peut être contesté, parce que la définition du mot reste dans l'ensemble celle de la norme écrite, et que la délimitation des phrases est la même. La norme est donc bâtarde en ce sens que des

critères empruntés à la forme écrite sont utilisés pour traiter la forme parlée[2]. Mais s'agissant d'un texte initialement connu sous forme écrite, pour lequel il n'y a pas de forme orale effectivement produite spontanément, il n'y avait pas moyen de procéder autrement. Simplement il n'est plus question, cette fois, de compter les ponctuations comme des mots, puisqu'elles n'apparaissent plus du tout à l'oral, et qu'il n'existe pas de moyen normé de transcrire l'intonation. Par ailleurs on a procédé autrement pour les nombres. Dans la transcription, on a compté pour un mot chacun des adjectifs numéraux utilisés, alors que pour la forme écrite, on a traité cette fois les nombres présentés sous forme de chiffres selon le principe graphique qui veut que toute suite continue constitue un mot unique.

Comme ici le nombre des données est plus modeste, on pourra les fournir. Le tableau 1 montre quelles sont les moyennes observées pour les textes transcrits. Le tableau 2 montre les paramètres correspondants pour la forme écrite des mêmes textes. On voit que si le nombre des mots par phrase ne bouge presque pas, le nombre d'éléments par mot est sensiblement différent. L'usage des digrammes et des lettres muettes augmente de plus d'une unité par mot le nombre moyen des signes nécessaires.

Quelles sont les corrélations observées ici ?

Pour la forme orale :

Corrélation linéaire $r = 0{,}94814$

Corrélation des rangs $R = 0{,}95152$

Pour la forme écrite :

Corrélation linéaire $r = 0{,}90851$

Corrélation des rangs $R = 0{,}89091$

Tableau 1.

Texte	Mots/Phrase	Phon/Mot
T3	15,3529	3,3978
T7	19,9388	3,4626
T26	20,8182	3,5735
T29	15,9149	3,3035
T36	21,6111	3,5874
T48	36,5833	3,9408
T53	12,1695	3,3593
T73	32,0476	3,7712
T80	28,0000	3,6752
T98	22,5532	3,6981
Moy.	19,5022	3,5488

Les textes (articles) portent des numéros de 1 à 103. Ils ont été sélectionnés par tirage au hasard parmi les 103 textes traités sous forme écrite. Les moyennes relatives à chaque texte ont été obtenues à partir du nombre total de phonèmes, de mots et de phrases du texte.

Si maintenant nous examinons les corrélations internes aux dix textes transcrits, nous trouvons 9 coefficients négatifs sur 10. Comme précédemment,

- à l'intérieur d'un même texte, la loi de Menzerath semble bien s'appliquer en ce sens que les phrases longues ont généralement des mots de longueur moyenne plus faible que les phrases courtes ;
- d'un texte à l'autre, au contraire, le type variable de texte (ce que j'ai interprété comme le degré de technicité du texte) aboutit au résultat contraire : les textes qui ont des phrases relativement longues usent aussi d'un vocabulaire composé en moyenne d'unités plus longues.

Tableau 2

Texte	Mots/Phrase	Lettres/Mot
T3	15,3529	4,6558
T7	19,8571	4,7595
T26	19,8788	4,7515
T29	15,9362	4,4446
T36	19,8333	4,7759
T48	36,0000	5,1181
T53	12,1695	4,5515
T73	31,6667	5,1278
T80	27,2188	4,8347
T98	21,9574	4,9331
Moy.	19,12329	4,77185

Voyez la légende du tableau 1. Voici les paramètres correspondants trouvés pour la forme écrite. Toutefois on a appliqué ici une norme dans laquelle chaque nombre constitue un seul "mot" et où les traits d'union sont négligés.

Je ne me prononcerai pas sur la valeur extrêmement forte des coefficients de corrélation qu'on vient d'obtenir. Certes, les précautions d'usage ont été prises pour que les dix articles soient représentatifs de l'ensemble des 103 articles traités précédemment. Mais je ne vois pas pourquoi un groupe de dix textes manifesterait en général une corrélation beaucoup plus forte qu'un ensemble plus important.

Les droites de régression de G. Altmann

Dans l'ouvrage collectif de 1989, G. Altmann propose (p. 16-21) une formule décrivant la distribution de la variable "nombre de mots par syntagme" en fonction de la variable "nombre de syntagmes par phrase" à partir d'un ensemble d'observations. J'ai simplement remplacé ici les deux variables, en calculant la distribution de la variable "nombre de lettres (de phonèmes) par mot" en fonction de la variable "nombre de mots par phrase".

G. Altmann part de l'idée que si, au lieu de travailler sur les grandeurs observées Y_i (ici nombre de lettres par mot) et X_i (ici nombre de mots par phrase), on opère sur leur logarithme, on peut se fonder sur l'idée plausible

2. L'autre conséquence du procédé est que, la ponctuation n'ayant pas été intégrée dans les transcriptions, on ne parlera plus ici du niveau du "groupe" interne à la phrase.

qu'il existe entre les deux une relation linéaire (du type $y' = ax' + b$). Il utilise la méthode des moindres carrés pour calculer la droite la mieux adaptée à la description de la distribution. Pour cela, il se sert des notations et formules suivantes : soit x_i un certain nombre de mots par phrase, y_{ij} l'observation d'un nombre de lettres ou de phonèmes par mot dans une des tailles de phrase. Si on désigne par n_i le nombre de phrases ayant x_i mots, et par n la somme de tous les n_i, on obtiendra le nombre moyen de lettres (phonèmes) par mot en calculant

$$\bar{y}_i = \sum_j \frac{y_{ij}}{n_i x_i} .$$

Par ailleurs[3], on note $Z_i = \ln x_i$; $y_i = \ln \bar{y}_i$; $A = \ln a$, qui est l'un des paramètres dans la formule $\ln \bar{y}_i = \ln a + b \ln x_i$. qu'on pourra récrire avec les autres notations : $y_i = A + bZ_i$. L'intérêt de cette formule est dans l'interprétabilité linguistique des deux paramètres introduits : A est identifié par Altmann avec la taille moyenne de la variable étudiée lorsque la variable contrôlée ne comporte qu'un seul constituant (dans notre cas, la longueur du mot dans la phrase d'un seul mot). Le paramètre b, quant à lui, définit la pente de la courbe : il dit avec quelle rapidité la longueur moyenne du mot va baisser (ou augmenter si b était positif) lorsque la phrase s'allonge.

On calculera

$$M_{ZZ} = \sum_i n_i Z_i^2 - \frac{\left(\sum_i n_i Z_i\right)^2}{n}$$

$$M_{ZY} = \sum_i n_i Z_i Y_i - \frac{\sum_i n_i Z_i \sum_i n_i Y_i}{n}$$

$$M_{YY} = \sum_i n_i Y_i^2 - \frac{\left(\sum_i n_i Y_i\right)^2}{n},$$

ce qui nous permet d'obtenir les coefficients attendus : si nous adoptons les notations suivantes :

$\bar{Y} = \frac{1}{n}\sum_i n_i Y_i$ et $\bar{Z} = \frac{1}{n}\sum_i n_i Z_i$, alors nous pouvons définir

$b = \frac{M_{ZY}}{M_{ZZ}}$ et $A = \bar{Y} - b\bar{Z}$, et $a = e^A$.

Nous pourrons de cette façon calculer les valeurs "théoriques" des moyennes y_i. Un test F nous permettra de voir si la pente définie par le paramètre b est significativement différent de 0, et si oui, s'il est positif ou négatif. S'il est négatif, le résultat viendra à l'appui de la loi de Menzerath ; s'il est positif, la tendance observée sera opposée. L'hypothèse nulle est que b n'est pas significativement différent de 0.

Les calculs ont été faits, sur les 103 textes écrits, de deux manières successivement : d'une part dans chacun des textes pris en lui-même, puis sur la réunion de tous les textes. Parmi les facteurs b observés pour les 103 textes individuels, nous en avons 66 qui sont négatifs, ce qui représente une proportion significativement supérieure à la moitié ; de plus, une vingtaine des tests où b est négatif franchissent le seuil classique de 5 %, ce que le simple hasard n'amènerait à peu près jamais, alors qu'on n'a que trois résultats positifs franchissant ce seuil, ce que le hasard, sur ce nombre d'alternatives, amène plus de trois fois sur quatre. Ces résultats indiquent que la loi de Menzerath semble s'appliquer, quoique de manière assez discrète.

Si nous soumettons au calcul l'ensemble des 4269 phrases totalisées par les 103 textes, nous obtenons b - 0.039, une valeur très faible, mais qui, au regard du nombre des données, est hautement significative : $F(1,80) = 40{,}087$, ce qui nous donne une probabilité de l'ordre 1 chance sur cent millions. Là aussi, la conclusion est la même. Cette fois, par conséquent, nous n'obtenons pas de résultats contradictoires selon que nous travaillons sur les textes séparés ou sur leur réunion. Mais remarquons que nous avons ici considéré toujours les nombres de mots par phrase et les nombres de lettres par mot, non les moyennes par texte. C'est ce qui fait la différence par rapport aux calculs de corrélation qui précèdent.

[3] On peut être choqué ici par le fait que le symbole y_i est utilisé pour une valeur très élaborée de la variable ; mais il vaut sans doute mieux réserver les symboles simples à des valeurs qui ont à être utilisées dans des formules particulièrement complexes.

Les mêmes calculs faits sur les transcriptions phonétiques donnent des résultats analogues, mais pas plus nets, cette fois, que ceux qu'on vient de voir. Sur l'ensemble des textes transcrits, nous avons
$b = 0{,}029$, et $F(1,56) = 5{,}201$
Ce résultat est significatif au seuil de 5 %. En faisant un test F pour chacun des dix textes, nous en obtenons trois qui sont significatifs, tous trois avec *b* négatif. Là aussi, la loi de Menzerath se trouve confirmée, toujours de manière aussi discrète.

"Loi" de Menzerath et "loi" d'Arens

Je mets ici des guillemets parce que les auteurs du livre d'Altmann & al. 1989 sont en désaccord entre eux sur le statut de la "loi de Menzerath", certains préférant parler de "règle" plutôt que de "loi". Une observation supplémentaire, faite en 1965 par H. Arens sur un ensemble de 117 textes allemands, a été étudiée par G. Altmann. Arens observe que plus les phrases de ses textes sont longues, plus les mots qui les composent sont longs eux aussi. G. Altmann (1983) en conclut qu'on a là un effet de la "transitivité" de la loi de Menzerath, en l'occurrence du fait que les phrases se décomposant en syntagmes ou en propositions, les syntagmes ou propositions en mots, des phrases plus longues exigent des propositions ou syntagmes plus courts, et les propositions ou syntagmes plus courts exigent des mots plus longs…
Je me permettrai ici de mettre en doute cette conclusion. Comme on l'a vu, nous avons obtenu une corrélation positive entre la longueur moyenne des phrases de nos textes et la longueur moyenne des mots de ces mêmes textes ; en étudiant au contraire les phrases et les mots à l'intérieur de chaque texte, nous constatons que généralement cette corrélation est plutôt négative, et que les phrases courtes ont des mots plus longs que les phrases plus longues.
Or précisément Arens travaille sur les moyennes observées sur les textes pris globalement, non sur les corrélations internes à chaque texte. Certes l'observation que nous avons faite ici est dérangeante en ce sens qu'elle parait contradictoire dans un premier temps ; mais dans un deuxième temps, on peut en proposer une interprétation cohérente : à l'intérieur d'un texte thématiquement et stylistiquement cohérent, la loi proprement linguistique s'applique, quoique fort discrètement ; dans la comparaison entre textes, ce sont les contraintes thématiques et les choix de présentation (plus ou moins grande technicité, par exemple) qui dominent. Malgré le choix, ici, d'un genre textuel relativement unitaire, ces dernières différences apparaissent de façon observable.

Conclusions provisoires

Résumons l'ensemble des résultats obtenus.
- Parmi les 103 articles examinés, ceux qui ont les phrases en moyenne les plus longues par le nombre des mots ont aussi des mots plus longs en moyenne que les textes dont les phrases sont plus courtes.
- À l'inverse, si nous examinons chaque texte en lui-même, les phrases les plus longues ont des mots en moyenne plus courts que les phrases plus courtes, dans près de trois textes sur quatre.
- Si nous calculons la droite de régression par la méthode des moindres carrés, le paramètre *b*, indiquant l'orientation globale de la pente, est négatif de façon nettement prédominante, ce qui conforte la loi de Menzerath.
- En opérant sur des transcriptions phonétiques de quelques-uns de ces textes, on trouve les mêmes résultats, et - soit hasard, soit nécessité linguistique - les calculs de corrélation donnent ici des résultats encore plus nets que dans le cas des textes écrits, alors que la droite de régression ne révèle pas plus nettement que dans les textes écrits la tendance observée.
- Le choix de la norme, c'est-à-dire essentiellement la définition du mot, intervient notablement dans le résultat ; en particulier la norme très pragmatique (au sens *non* technique du terme) pratiquée dans la base Frantext est de nature à fausser les conclusions et à produire, sur le point qui nous a intéressés ici, un artefact sans véritable signification linguistique ; il y a lieu d'écarter du compte des mots tous les signes non alphabétiques ; le traitement des chiffres n'est pas sans incidence non plus.
On peut tirer de l'ensemble de ces constatations plusieurs conclusions méthodologiques qui sont de nature à autoriser un regard nouveau sur la loi de Menzerath, et par ailleurs elles nous amènent à poser d'autres questions.
Tout d'abord il n'est nullement indifférent qu'on travaille sur un inventaire de formes, ou sur un texte, ou encore sur un ensemble de textes. Les conditions et la population sur lesquelles on a obtenu tel ou tel résultat doivent être indiquées avec la plus grande précision ; l'interprétation, ou l'interprétabilité, de ces résultats peut en dépendre. Menzerath lui-même n'examinait que le lexique, et jugeait même que la considération des fréquences n'avait aucun rôle structurel (sans s'appesantir sur ce qu'il entendait précisément par "rôle structurel"). G. Altmann (1989:51, §5.3.1.) lui-même n'accorde pas une "grande pertinence" à la prise en compte des fréquences. Or il suffit de considérer le niveau d'analyse où se plaçait P. Menzerath pour trouver ce dernier point de vue hautement contestable : la grande majorité des unités (mots) de très grande fréquence

sont des monosyllabes. Les monosyllabes allemands peuvent comporter de 1 à 6 phonèmes sans difficulté (*eh', Brunst, stracks, Splint*) ; mais les monosyllabes les plus courants n'ont jamais que deux ou trois phonèmes, c'est-à-dire une ou deux consonnes : ce sont des mots tels que les articles *der, die, das, den, dem*, les prépositions *an, bei, zu, um, von, und*. La prise en compte des fréquences a donc toutes les chances de faire baisser la longueur moyenne du monosyllabe dans les textes par rapport à sa longueur moyenne dans le lexique des formes. Dans les faits, les résultats obtenus au niveau du mot sur des inventaires paraissent plus nets que ceux qui ressortent de l'analyse de textes. Il n'est pas impossible que ce contraste lui-même appelle une interprétation en termes d'économie de moyens.

Ensuite la nature de la formule obtenue par Altmann implique que le rôle essentiel est joué dans les faits par les toutes premières lignes du tableau des répartitions ; dans le cas des mots et de leurs syllabes, par les mots de une ou deux syllabes, en fait. On peut remarquer dans nos propres données aussi que les phrases de 1 à 4 mots ont des mots en moyenne beaucoup plus longs que les phrases plus longues, et que la longueur moyenne diminue fortement d'une de ces toutes premières classes à la suivante ; après cela, les irrégularités se multiplient, et nos phrases de 5 mots ont une longueur moyenne du mot plus faible que les phrases de 90 mots ou plus. On peut alors se demander si la loi à considérer ne serait pas plutôt une loi relative aux unités les plus courtes seulement. Il est après tout concevable que les phrases les plus courtes aient besoin d'atteindre un volume critique nécessaire pour être correctement perçues. Je ne prétends pas énoncer ici une telle loi, mais il me semble que la question mérite d'être posée.

La loi de Menzerath peut bel et bien être une loi structurelle des langues, mais en même temps il faut se garder de vouloir l'appliquer sans précaution à la comparaison des textes d'une même langue. Gabriel Altmann avait mis en garde lui-même, dans Altmann & al. 1989:8-14, contre les "facteurs perturbants" qui peuvent venir troubler les résultats. Les contraintes thématiques ou autres qui s'exercent dans le discours peuvent amener certains textes à être plus "compacts" que d'autres, à user de constituants plus longs. Par la même occasion, si nous comparons des langues typologiquement peu éloignées les unes des autres, le choix des faits linguistiques particuliers sur lesquels nous travaillons n'est pas indifférent.

Il faut sans doute se garder de raisonner en termes de niveaux d'analyse strictement cloisonnés ; dans trois niveaux, nommés N1, N2 et N3 du plus global au plus restreint (p.ex. N1 : phrase ; N2 : syntagme ; N3 : mot), Altmann imaginait, on l'a vu, que si les unités de niveau N1 devenaient plus grandes, les unités du niveau N2 devaient devenir plus petites, et donc les unités du niveau N3 plus grandes ; quoiqu'il ait été très prudent dans l'expression de cette "transitivité", il semble bien, d'après ce qui vient d'être dit, qu'il ne faille pas l'envisager du tout - cependant d'autres recherches seraient nécessaires pour en avoir le cœur net. Si, à la différence de Menzerath, nous choisissons de travailler sur du discours, et non sur un inventaire d'unités disponibles, nous sommes sur un terrain beaucoup plus mouvant, puisque rien ne nous garantira que les textes que nous sélectionnons dans les langues à comparer soient effectivement comparables. A l'intérieur d'une même langue, il sera difficile de garantir objectivement que les critères de distinction qui opposent deux textes ou deux ensembles de textes aient été correctement et exhaustivement définis. Il y a là un domaine d'étude qui est encore ouvert : définir les conditions dans lesquelles la loi de Menzerath peut être raisonnablement supposée s'appliquer. Nous avons vu que cela semble bien être le cas à l'intérieur d'un texte homogène ; ce n'est pas le cas, en revanche, dans la comparaison de textes qui font pourtant tous partie d'un "genre" relativement bien délimité, comme celui de la prose d'un quotidien d'information intellectuel. Rien ne nous dit à quels résultats nous sommes en droit de nous attendre si nous comparons différents textes d'œuvres dramatiques, ou des textes de romans, ou des œuvres poétiques - trois genres que l'intuition peut nous faire supposer intrinsèquement bien plus divers que celui que nous avons étudié ici. Nous ne savons pas non plus comment se comportent de ce point de vue des textes d'un même auteur. La prose *orale* spontanée suit-elle des règles analogues à celles que nous avons observées dans des textes délibérément composés pour rester sous forme écrite ? Il serait nécessaire aussi de mettre la loi de Menzerath en relation avec la tendance à l'isochronie des segments successifs, tendance relevée depuis plus d'un siècle (Sievers[4] 1901:264) et dont l'étude n'est guère avancée. Il s'agit là d'un travail nouveau, dans lequel les segments successifs ne se confondent nullement avec un niveau syntagmatique. Ces questions et sans doute quelques autres pourraient être soulevées, et je serais comblé si elles suscitaient la curiosité des chercheurs.

4. Altmann (1989:60) cite ce passage ; mais là où il écrit *Sprechakte* ("actes de langage"), il faut lire *Sprechtakte* ("groupes rythmiques"). C'est intentionnellement que je ne mentionne ici que l'édition de 1901, et non l'édition initiale de 1876, où ce passage ne figure pas.

Références bibliographiques

[ALTMANN 1989] Altmann, Gabriel & Schwibbe, Michael H., *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, mit Beiträgen von Werner Kaumanns, Reinhard Köhler und Joachim Wilde, Hildesheim-Zürich-New York, Georg Olms, 1989.
[ALTMANN 1983] Altmann, Gabriel "H. Arens' «verborgene Ordnung» und das Menzerathsche Gesetz", *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik*, ed. Faust M., Harweg R. & Lehfeldt W. 1983:31-39.
[ARENS 1965] Arens, H. *Verborgene Ordnung*, Düsseldorf, Schwann, 1965.
FRANTEXT, base de données textuelles, http://atilf.atilf.fr/frantext.htm (sur abonnement).
[MENZERATH 1954] Menzerath, Paul, *Die Architektonik des deutschen Wortschatzes*, Bonn-Hannover-Stuttgart, Ferd. Dümmler, 1954.
[MULLER 1973] Muller, Charles, *Initiation aux méthodes de la statistique linguistique*, Paris, Hachette, 1973. (Réimprimé chez Champion en 1992).
[ROUKK 2003] Roukk, Maria, "The Menzerath-Altmann law in Russian texts (sentence level)", comm. au *4. Trierer Kolloquium zur Quantitativen Linguistik*, 16-18 octobre 2003, Trèves (à par., sans doute dans *Journal of Quantitative Linguistics*).
[SIEVERS 1901] Sievers, E. *Grundzüge der Phonetik*, Leipzig, Breitkopf & Härtel, 1901.

La série *Glottometrika* publiée chez Brockmeyer à Bochum, et qui entre dans la collection *Quantitative Linguistics* dirigée par G. Altmann et R. Grotjahn, contient dans ses numéros 2 (1980), 4 (1982), 5 (1983) et 6 (1984) toute une série d'articles concernant la loi de Menzerath, mais qui n'ont pas été nommément cités ici. Ils sont presque tous en allemand. Dans le numéro 6, Reinhard Köhler ("Zur Interpretation des Menzerathschen Gesetzes", p. 177-183) fait la synthèse de cette série d'articles en en proposant une interprétation globale.