Evénements langagiers rares et acquisition du langage

Christophe Parisse

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

Evénements langagiers rares et acquisition du langage

Christophe Parisse

Inserm/Modyco, CNRS / Université Paris Ouest Nanterre
cparisse@u-paris10.fr

1 Résumé

La loi de Zipf-Mandelbrot décrit la répartition des événements langagiers comme suivant une loi de puissance. Une des propriétés de cette loi est que les éléments les plus rares sont chacun uniques, mais que globalement ils sont très diversifiés et très nombreux. Cette loi est le plus souvent considérée seulement du point de vue de l'écrit et de la compréhension du langage.

Théoriquement elle devrait aussi s'appliquer aux corpus oraux, y compris au langage produit par une seule personne. Si c'était le cas, alors la conséquence surprenante est qu'un locuteur produit un grand nombre de termes une seule fois (ou peu de fois) au cours de sa vie, et qu'une grande partie du vocabulaire acquis dans l'enfance ou utilisé à l'âge adulte n'est basé que sur quelques exemplaires.

L'article teste l'existence des caractéristiques de la loi de Zipf sur une série de corpus enfant-adulte, y compris un corpus dit dense de 2 millions de mots. Ne disposant pas de corpus de « taille humaine » (plusieurs années de production ou de compréhension du langage), la comparaison avec des corpus écrits est également faite.

Il ressort que rien ne permet d'infirmer l'application de la loi de Zipf aux corpus de langage oral, et en particulier les conséquences pour les items rares. Ces conséquences, prises au sérieux, ont une répercussion importante sur le système langagier. Une discussion de ces conséquences et de leurs implications pour la recherche en linguistique est présentée en conclusion.

2 Introduction

La distribution des mots dans un corpus de langage suit une loi d'abord décrite par Zipf (1949) puis raffinée par Mandelbrot (1962). Le principe de base de cette distribution est de suivre une loi de puissance, c'est-à-dire que lorsque les éléments sont ordonnés en fonction de leur fréquence, ils suivent une loi exponentielle (chaque élément est *n* fois plus fréquent que l'élément qui le suit, avec *n* qui varie suivant la pente de la distribution). On peut dire de la distribution de Zipf-Mandelbrot qu'elle suit une loi de Pareto (Newman, 2005), c'est-à-dire que les 20% d'éléments les plus fréquents représentent 80% des données en tout. En d'autres termes, les quelques éléments les plus fréquents représentent une très grande partie des données et inversement, les éléments rares, très nombreux, ne représentent qu'une faible partie des données.

Une autre image de la distribution de Zipf est de parler de distribution à queue longue. Cette queue fait ici référence aux éléments de faible fréquence. Chaque élément peu fréquent a un impact faible sur l'ensemble de la distribution mais la somme de tous les éléments rares représentent une part importante de la distribution, ou même une large majorité des données selon la valeur du paramètre qui définit la distribution. Une application de cette propriété est par exemple intéressante en économie. Le marché potentiel que représentent de multiples ventes de petites valeurs peut correspondre à un marché global de grande importance.

De la même façon en langage, l'ensemble des items ayant un faible nombre d'occurrences chacun peut représenter globalement une part importante de l'ensemble de notre expérience langagière. Une différence importante entre l'image de l'économie et celle du langage est celle des généralisations que l'on peut tirer des éléments rares. Si tous les événements rares représentent une situation de même type, chaque fois différente mais présentant toutes un point commun applicable globalement, alors les événements rares sont intéressants à traiter et peuvent avoir un effet considérable sur l'ensemble du système. Mais si chaque item est spécifique alors la généralisation d'un profit global est plus difficile à mettre en place.

Même si la loi de Zipf est largement acceptée dans toutes les descriptions de corpus de langage, elle implique des propriétés qui ne sont pas nécessairement intuitives dans toute situation de langage. Par exemple, la loi de Zipf indique que de très nombreux items n'ont qu'une seule occurrence, que *n* fois moins d'items ont deux occurrences et ainsi de suite. Cette abondance d'apaxes ou de quasi-apaxes est facile à imaginer pour des corpus artificiellement rassemblés par l'association de sources multiples, mais elle n'est pas intuitive pour le corpus formé de la production d'un seul individu. Ainsi, est-il intuitif d'imaginer que nous n'utilisons dans notre propre

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801339

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

vie un grand nombre de mots qu'une seule fois, un nombre à peine plus faible deux fois (dans toute notre vie) et ainsi de suite ? Il est assez intuitif d'imaginer que quelques mots sont produits (ou perçus) un grand nombre de fois, mais l'inverse est moins vrai. Par ailleurs, si cela est vrai, est-ce que ces évènements linguistiques rares (produits ou perçus) forment des items globalement généralisables ou chaque item est-il unique et totalement différent des autres ?

Le but de cet article est de poser les conditions d'une analyse et d'un traitement des événements langagiers rares. Pour cela, l'article cherche à répondre aux questions suivantes :

- Peut-on démontrer, en dépit du fait que l'on ne dispose aujourd'hui d'aucun corpus transcrit décrivant exhaustivement les productions d'une personne, que le corpus formé par le matériel produit (et entendu) par une seule personne unique respecte la loi de Zipf-Mandelbrot ?
- Est-il possible de décrire précisément les propriétés de distribution des événements rares, en particulier en langage oral ?

Les réponses à ces questions entraineront une discussion qui portera sur les points suivants :

- Quelles sont les conséquences de l'existence de ces propriétés sur les théories linguistiques ?
- Les théories actuelles sont-elles capables d'expliquer les phénomènes langagiers pour toutes les échelles de fréquence ?
- Comment peuvent-elles évoluer pour expliquer le fonctionnement langagier dans les situations d'événements rares ?

3 Evaluation de la taille du corpus oral

Le propre d'un corpus de langage produit par une même personne est de contenir les transcriptions de toutes les interactions de cette personne avec autrui ou les interactions qu'elle peut avoir entendues sans y participer, et ne contenir que ce matériel. Il est raisonnable de penser qu'une personne passe autour de 10 heures par jour à parler ou à entendre des personnes parler. Cette valeur peut probablement varier de manière importante selon les personnes, les situations personnelles, les styles, etc. Tomasello et Stahl (2004) se basent par exemple sur cette valeur de 10 heures. Ce n'est qu'une valeur indicative dans notre contexte. Les valeurs moyennes extrêmes, hors cas pathologiques ou très spécifiques, ne vont probablement pas en deçà et au delà de 2 heures à 18 heures par jour. Toutes valeurs dans cette fourchette ne modifient pas les résultats présentés ci-après, car on considère des corpus variant en taille d'un facteur 10. Lorsque cela paraîtra nécessaire, nous préciserons l'intervalle des valeurs extrêmes sur lesquelles nous pouvons faire nos estimations. Dans les autres cas, nous partirons d'une valeur moyenne de 10 heures par jour.

Sur cette base, un corpus contenant toutes les interactions d'une même personne contiendra environ 365 jours fois 10 heures, c'est-à-dire 3650 heures de corpus par an. Si cette personne est un adulte d'une quarantaine d'années, on est face à un corpus de 40 fois 3650 heures, c'est-à-dire 14 600 heures. Le débit horaire d'un adulte est également variable. Pour prendre des exemples concrets, le Corpus du Français Parlé Parisien (CFPP2000 : http://cfpp2000.univ-paris3.fr/) présente un débit horaire moyen de 6060 mots, le corpus adulte de TCOF (Atilf : http://www.cnrtl.fr/corpus/tcof/) un débit moyen horaire de 4930 mots. Le corpus d'interaction avec l'enfant de TCOF présente un débit moyen horaire de 2650 mots (adultes et enfants).

Les corpus d'interaction avec de très jeunes enfants présentent des débits moindres, de 420 mots par heure pour les enfants du corpus Paris (Prismes/Modyco/Childes : http://colaje.scicog.fr/index.php/corpus), 1480 pour les adultes. Les valeurs sont de 620 mots pour l'enfant et 2380 pour le corpus de Thomas (enfant anglophone/Childes : http://childes.psy.cmu.edu/data/Eng-UK/). Les corpus CFPP2000 et TCOF correspondent à la somme de plusieurs interlocuteurs adultes et correspondent logiquement à environ le double des débits des adultes dans les corpus dont l'un des interlocuteurs principaux est un enfant et les autres sont des adultes. Tous ces corpus sont des situations de conversation en général soutenue, ce qui n'est pas nécessairement une situation naturelle permanente. En conséquence on peut imaginer que le débit d'un adulte horaire varie entre 1000 et 4000 mots, celui d'un jeune enfant se rapprochant plus d'une moyenne de 500 mots.

On peut donc faire une estimation de l'ordre de grandeur du corpus produit et entendu par un adulte et un enfant. Pour les adultes, on considèrera qu'ils entendent 2 fois plus de matériel sonore qu'ils ne produisent. Dans ces conditions, ils produisent en moyenne 1000 mots par heure et entendent 2000 mots (sur la base d'un débit horaire moyen global en interaction de 3000 mots). Pour les enfants, on considéra qu'ils entendent 1500 mots (légèrement moins qu'un adulte) et qu'ils produisent 500 mots par heure.

Sur ces bases on peut estimer qu'un enfant de trois ans a entendu 3 ans x 3650 heures x 1500 mots, c'est-à-dire 16 425 000 mots (16,5 millions). Un adulte de 40 ans a entendu 40 x 3650 x 2000, c'est-à-dire environ 292

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

millions de mots (chiffre qui pourrait être ramené à 282 millions en tenant compte d'une réception plus faible pendant les 5 premières années de sa vie). Sur les mêmes bases, un enfant de 3 ans aura produit environ 3 x 3650 x 500 mots, c'est-à-dire 5,5 millions de mots, et un adulte de 40 ans aura produit 141 millions de mots ! Pour faciliter les comparaisons avec toutes tailles de corpus on peut aussi retenir les chiffres estimés annuels : pour l'enfant, 1,8 millions de mots produits et 5,5 millions de mots entendus ; pour l'adulte, 3,6 millions de mots produits et 7,3 millions de mots entendus.

4 Corpus de langage oral et écrit

Sur la base des chiffres précédents, on peut évaluer la différence existant entre les corpus de langage oral dont nous disposons et la taille de corpus qui couvriraient effectivement toutes les productions d'une personne. Nous avons sélectionné 6 corpus pour effectuer nos évaluations :

- Thomas (source CHILDES) : corpus d'un seul enfant enregistré sur 2 ans environ - http://childes.psy.cmu.edu/data/Eng-UK/
- Manchester (source CHILDES) : corpus de 12 enfants enregistrés sur 1 an - http://childes.psy.cmu.edu/data/Eng-UK-MOR/
- Paris (source Prismes/Modyco/CHILDES) : corpus de 4 enfants enregistrés sur 5 ans- http://childes.psy.cmu.edu/data/Romance/
- Lyon (source DDL/CHILDES) : corpus de 4 enfants enregistrés sur 2 ans - http://childes.psy.cmu.edu/data/Romance/
- OANC (source OAC) : corpus anglais de langage écrit et oral - http://www.americannationalcorpus.org/
- Gutenberg (source Gutenberg) : ensemble de textes écrits en langue anglaise principalement - http://www.gutenberg.org/

Cette sélection comprend Thomas, le corpus le plus dense librement disponible en langue anglaise (Lieven, Salomo & Tomasello, 2009). Les enregistrements de ce corpus ont été réalisés 1 heure par jour et 5 jours par semaine pendant une période allant pour l'âge de Thomas, de l'âge de 2 ans à l'âge de 3 ans et 3 mois. Le corpus se prolonge de manière moins dense (5 heures d'enregistrement par mois) pendant une année et demie de plus. Ce corpus de langue anglaise peut être comparé au corpus de langue anglaise Manchester (Theakston, Lieven, Pine & Rowland, 2001) présentant presque la même taille totale mais comportant le suivi de 12 enfants (au lieu de 1) avec des enregistrements effectués 2 fois par mois sur une période de un an (âge des enfants : 1 an 10 mois à 2 ans 10 mois).

On ne dispose pas de corpus dense en français, mais on peut constituer un corpus assez proche du corpus Manchester en réunissant les deux corpus Paris (Morgenstern & Parisse, 2012) et Lyon (Demuth & Tremblay, 2008).

Le corpus Thomas adulte ne représente qu'un huitième de la taille que ferait un corpus adulte correspondant à ce qu'un enfant de trois ans a entendu au cours de sa vie. S'il n'est pas possible de trouver aujourd'hui un corpus libre de droit et de la taille requise, on peut s'en rapprocher en utilisant plusieurs corpus : Thomas, Manchester et la base OANC (Open American National Corpus), dans sa partie orale, avec un total de 7,8 millions de mots, se rapprochent des 16 millions de mots qui serait la taille idéale.

Pour obtenir des corpus équivalents aux tailles atteintes pour la durée d'une vie d'un enfant ou d'un adulte, il faut (dans les corpus libres) utiliser des corpus de langue écrite. Ces corpus de langue écrite n'ayant pas nécessairement les mêmes propriétés que les corpus de langue orale, on réalisera d'abord une comparaison à taille égale avec des corpus oraux avant de tirer des conclusions pour des corpus de plus grande taille.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

Tableau 1 : Description des corpus

Corpus	Langue	Nb interlocuteurs	Nb occ. Enfant	Nb occ. Adulte(s)
Thomas	Anglais	1+1	510 000	2 000 000
Manchester	Anglais	12+12	530 000	1 450 000
Paris+Lyon	Français	8+8	302 000	1 271 000
OANC oral	Anglais	nombreux		4 290 000
OANC écrit	Anglais	nombreux		19 400 000
Gutenberg total	Anglais	nombreux		197 000 000

5 Description statistique des corpus existants

Le volume des corpus et la manière dont se répartissent les fréquences de mots qui les composent sont modélisés par les lois de Zipf (1949) et Zipf-Mandelbrot (1962). Ces deux lois ont été utilisées dans le package statistique zipfR (bibliothèque du langage R) de Everts et Baroni (2007) qui permet de calculer automatiquement les paramètres qui décrivent au mieux les données.

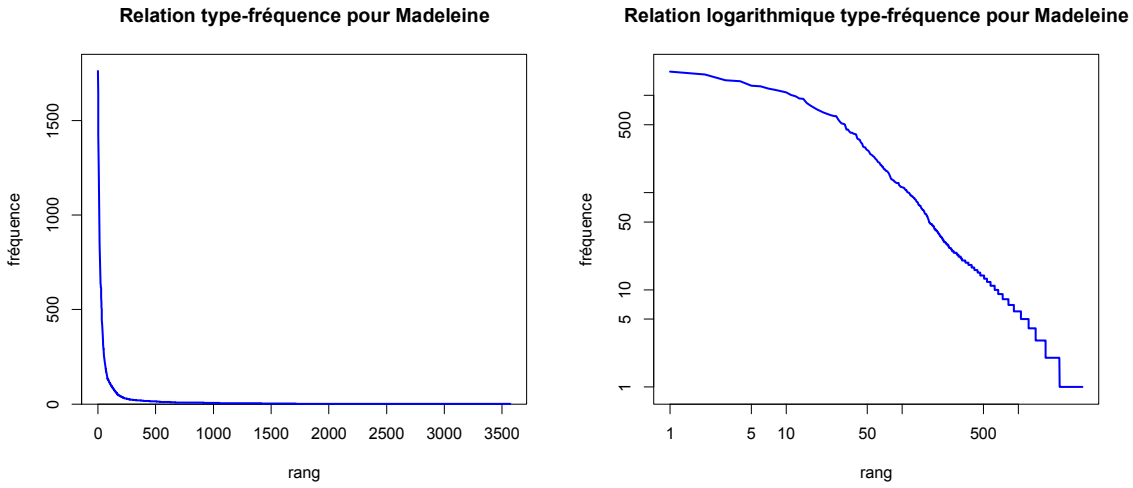
Le package se base sur des mots triés de manière à générer une table de fréquence : à chaque mot correspond son nombre total d'occurrences. Cette forme de représentation correspond à la forme classique des courbes de Zipf-Mandelbrot. Le mot le plus fréquent est α fois plus fréquent que le second mot le plus fréquent, le second α fois plus fréquent que le troisième plus fréquent, et ainsi de suite. Il s'agit donc d'une progression géométrique et α correspond au principal paramètre de la loi de Zipf-Mandelbrot. La loi de Zipf-Mandelbrot s'exprime de la manière suivante :

$f(\text{mot}) = C / (r(\text{mot}) + b)^{\alpha}$ où $f()$ correspond aux fréquences théoriques, $r()$ aux rangs des mots et α, b sont des paramètres. C n'est pas exactement un paramètre mais une constante qui dépend de la fréquence de l'élément le plus fréquent (et donc de la taille du lexique).

La représentation en fréquences/rangs est présentée à gauche dans la figure 1 et correspond aux résultats du corpus de Madeleine, partie correspondant à l'ensemble de la production de l'enfant. Il n'est pas facile de « zoomer » sur une partie de cette représentation car les hautes valeurs et les basses valeurs sont très « étalées », chacune dans des axes différents. La partie droite de la figure correspond à la même représentation mais avec des valeurs logarithmiques. La quasi-linéarité de cette courbe exprime bien le caractère géométrique de la loi de Zipf-Mandelbrot.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

Figure 1 : Relations fréquences-rangs des mots produits par Madeleine (corpus Paris)

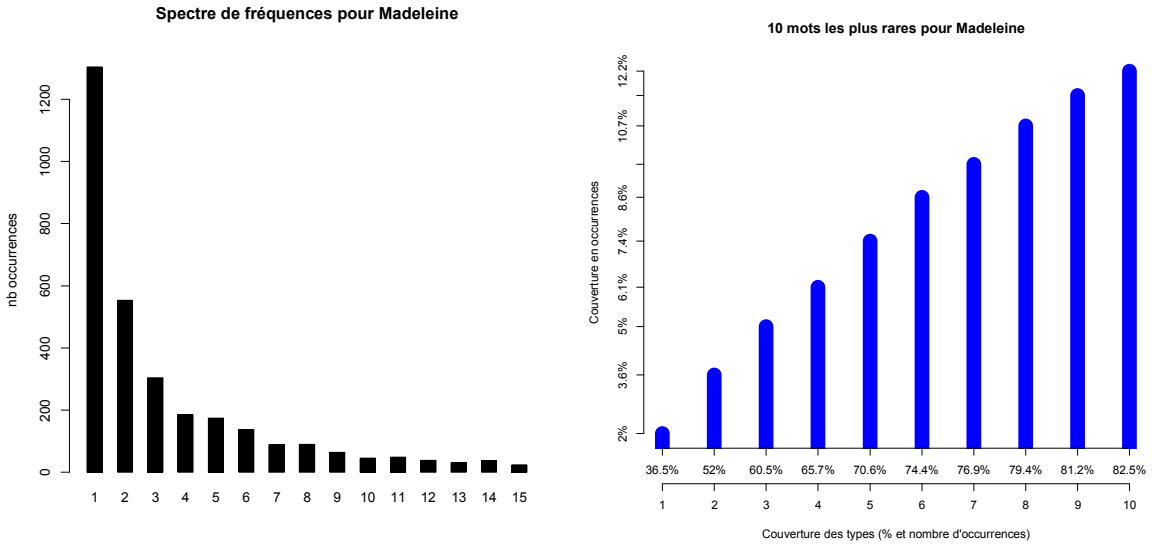


Une représentation plus intéressante des courbes de Zipf est la forme dite de « spectres de fréquences ». Dans ce cas, on calcule pour chaque point du spectre le nombre de types correspondant à un mot et le nombre d'occurrences. Par exemple, dans le corpus de Madeleine ci-dessus, le mot le plus fréquent (*c'est*) correspond à 1762 occurrences. Les 146^{ième} et 147^{ième} mots les plus fréquents (*doudou* et *prendre*) ont chacun 69 occurrences. En continuant de descendre dans l'échelle de fréquence, on trouve par exemple 49 mots de 11 occurrences, 185 mots de 4 occurrences et enfin 1304 mots de une seule occurrence. Cette suite de valeurs types/nombres d'occurrences correspond au spectre représenté de manière schématique par zipfR sur la partie gauche de la figure 2. Dans cette représentation, l'axe des y correspond aux fréquences des items, l'axe des x fournit une image globale des éléments, allant des plus rares à gauche aux plus fréquents à droite. Cette forme de représentation permet de présenter une image globale du spectre de fréquence. La dernière représentation qui est utile pour étudier les évènements les plus rares est présentée sur la partie droite de la figure 2. Ici seuls les éléments les plus rares sont représentés. L'axe des x correspond aux pourcentages de types (par rapport à l'ensemble des types du corpus) correspondant aux éléments uniques (à gauche), puis aux éléments répétés 2 à 20 fois dans le corpus. L'axe des y correspond aux pourcentages d'occurrences des mêmes éléments par rapport à l'ensemble des occurrences du corpus. Ce graphique dit « événements rares » permet de visualiser le comportement des mots produits le moins fréquemment dans un corpus.

Pour comparer la répartition des fréquences des mots rares et leur représentativité dans les corpus, il est plus simple de présenter les données sous forme de tableau. Pour se focaliser sur les éléments qui nous intéressent le plus dans ce travail et limiter la taille des données à visualiser, le tableau 2 présente les résultats en pourcentages cumulés de représentation des types et des occurrences des seuls mots figurant entre 1 et 5 fois dans un corpus. Les 6 corpus décrits dans le tableau 1 sont analysés. Pour les corpus contenant plusieurs locuteurs identifiés (Manchester et Paris+Lyon), deux lignes décrivent le corpus : une qui correspond aux valeurs moyennes pour un locuteur unique, l'autre qui correspond à la somme du corpus. Les corpus enfant et adulte sont séparés et considérés comme des corpus indépendants.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

Figure 2 : Représentation spectrale des nombres d'occurrences pour Madeleine (corpus Paris)



Les résultats présentés dans le tableau 2 montrent une nette différence entre les corpus de langue orale et écrite. Les corpus écrits tendent à avoir plus de mots rares que les corpus oraux, que ce soit pour les mots apparaissant 1, 2, 3, 4 ou 5 fois. Le coefficient α de la loi Zipf-Mandelbrot est également plus fort pour les corpus écrits, à l'exception du cas du corpus des enfants français, vu enfant par enfant « PL (moy) ». Les valeurs d'α pour ce corpus sont confirmées pour tous les enfants, y compris s'ils viennent de différents corpus et donc différents transcripteurs (corpus de Paris et de Lyon réunis). Tous les autres corpus oraux présentent des valeurs uniformes, pour le coefficient α comme pour le nombre de mots de faible nombre d'occurrence. Seul le corpus des enfants de Manchester présente des valeurs plus faibles, aussi bien calculées en moyenne qu'en rassemblant tout le corpus en un seul fichier.

Les résultats montrent que, pour tous les corpus, le nombre de mots apparaissant très rarement dans les corpus est, en types de mots par rapport à l'ensemble du corpus, très important. Pour les corpus oraux, les mots n'apparaissant qu'une seule fois représentent de 29,4% à 42,7% des mots des corpus (valeur minimale pour Manchester enfant en moyenne, valeur maximale pour Thomas enfant). Les mots apparaissant de 1 à 5 fois représentent de 59,8% à 72,7% des mots des corpus (valeur minimale pour Manchester enfant corpus complet, valeur maximale pour Paris-Lyon enfant en moyenne). Le très gros corpus OANC de langage oral (4,2 millions de mots), présentant de nombreux locuteurs différents, se comporte comme le corpus de Thomas enfant (500 000 mots) et Thomas adulte (2 millions de mots), ne comprenant tous deux qu'un unique locuteur. Les corpus écrits ont des tendances plus fortes, 56% de mots apparaissant une seule fois et 81% de mots apparaissant au maximum 5 fois. Ces deux corpus écrits qui présentent des tailles 5 à 50 fois plus importantes que les corpus oraux confirment donc les tendances en les amplifiant. Ces corpus sont exactement de la taille attendue pour couvrir toutes les productions d'une personne. Ils sont donc très instructifs et forment un test en grandeur nature. Toutefois il ne s'agit pas d'oral. Pour de vraies évaluations, il faudrait faire des interpolations de corpus plus petits ou créer de gros corpus artificiels en mélangeant des corpus d'origine diverse.

6 Modification de la taille des corpus

Ne disposant pas de corpus de taille adéquate pour évaluer précisément le comportement de grands corpus oraux vis à vis des mots les plus rares, on peut procéder par approximation et interpolations de trois manières :

- Les grands corpus oraux et écrits sont-ils de même caractéristiques que des corpus plus petits extraits des mêmes corpus ?

- Si oui, alors les petits corpus écrits sont-ils proches des grands corpus oraux ?

Pour juger si les corpus de même type mais étendu d'un facteur de 10 ont les mêmes propriétés, tous les corpus figurant dans le tableau 2 ont été divisés en 10 parties égales, puis les mêmes analyses ont été faites.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

Pour comparer les valeurs obtenues avec les valeurs initiales pour tous les corpus, la moyenne des valeurs a été calculée pour 6 mesures : pourcentage de mots n'ayant que 1, 2, 3, 4 ou 5 occurrences, valeurs du coefficient α de la loi Zipf-Mandelbrot. On a ensuite comparé cette moyenne aux valeurs des corpus complets. La présentation des valeurs totales en comparaison des valeurs moyennes est présentée dans le tableau 3.

Lorsque l'on compare les colonnes « Tout » et « 1/10 » correspondant aux corpus complets d'une part, et divisés par 10 d'autre part, on constate qu'à l'exception du corpus de Thomas pour l'enfant et l'adulte, du corpus Manchester complet adulte et du corpus OANC écrit, il y a une différence claire entre les deux tailles de corpus pour tous les paramètres (les items entre parenthèses dans le tableau 3 correspondent à des différences non-significatives, calculées à l'aide d'un t-test de Student, avec un seuil fixé à 0,01).

Il est difficile d'émettre un avis définitif à partir des seuls corpus de cette étude, mais il semble que pour les corpus oraux la réduction de la taille du corpus amène à une plus grande importance des mots rares en proportion du nombre de mots différents dans un corpus. Cette caractéristique est assez fiable car elle se vérifie pour tous les corpus oraux, sauf trois corpus où on ne trouve pas de différence. Ces trois corpus sont les plus volumineux. Pour les corpus écrits, on a plutôt l'inverse, une diminution de la proportion de mots rares dans le lexique. Mais les différences sont faibles et pas toujours significatives. On peut donc avoir l'impression que l'importance des mots rares tend à se stabiliser pour les corpus oraux lorsque les corpus deviennent très grands, tandis que l'inverse se produirait pour les corpus écrits.

L'opération de réduction de taille des corpus permet aussi d'obtenir des corpus écrits de taille comparable aux corpus oraux, ou d'autres corpus écrits. Ainsi, le corpus OANC écrit divisé par 10 a environ la même taille que le corpus adulte de Thomas et de Manchester au complet. On constate que la valeur du coefficient α de la loi Zipf-Mandelbrot est de 0,625 pour OANC complet, de 0,663 en moyenne (ET=0.034) pour OANC/10, de 0,485 pour Thomas, de 0,452 pour Manchester. Les valeurs du pourcentage de mots apparaissant une seule fois est de 59% pour OANC, de 55% en moyenne (ET=0.027) pour OANC/10, 40% pour Thomas et 36% pour Manchester. On peut aussi voir que le corpus Gutenberg divisé par 10 a des valeurs proches de celles du corpus OANC écrit : 0,605 pour le coefficient α de la loi Zipf-Mandelbrot et 53% pour les mots apparaissant une fois.

On constate que l'importance plus grande des mots rares dans les corpus de langage écrit que de langage oral reste vraie même pour les sous-parties de corpus, en dépit du rapprochement des valeurs dû à l'augmentation de l'importance des mots rares dans les corpus oraux réduits.

7 Propriétés des corpus de langage oral

Les résultats précédents ne forment pas une démonstration complète des caractéristiques fréquentielles des textes de langage oral pour les mots rares, car on ne dispose pas de corpus complet aujourd'hui, c'est-à-dire de corpus couvrant toutes les productions d'un interlocuteur.

On peut néanmoins constater sur des corpus représentant déjà près d'un dixième d'un corpus complet que les caractéristiques de la loi de Zipf semblent s'appliquer parfaitement. Il y a certes une différence nette entre les corpus oraux et écrits : les corpus écrits présentent jusqu'à 50% de plus de variété dans les mots n'apparaissant qu'une fois dans les corpus. Ce ratio n'est plus que de 10-20% lorsque l'on considère les mots apparaissant 5 fois dans le corpus. Cette différence peut s'expliquer par la diversité des sources écrites opposées à l'unicité des autres sources orales : des corpus d'acquisition du langage tous obtenus dans des conditions similaires.

L'argument de la diversité est quelque peu mis en défaut par les propriétés du corpus oral OANC, issu des conversations de 500 locuteurs différents (http://www.anc.org/data/oanc/contents/#switchboard). Les propriétés de ce corpus sont en effet très proches de celle du corpus Thomas, d'une taille proche et ne contenant que deux locuteurs.

Il est donc possible que le langage oral présente une diversité légèrement plus faible que celle du langage écrit. Néanmoins, cette diversité des mots les moins fréquents reste énorme :

- 30 à 40% du lexique n'est attesté qu'une seule fois, en compréhension comme en production ;
- 15% à 20% du lexique est attesté deux fois ;
- 20% du lexique est attesté 3, 4 ou 5 fois.

On a donc d'une façon générale environ 3 mots sur 4 dans notre lexique en réception et en production qui ne sont produits pas plus de 5 fois dans un corpus. Dans l'exemple de Thomas, on couvre environ 10% du temps de parole : « peut-on espérer que les 90% restant réutilisent uniquement les mêmes mots ? » Cela semble peut probable car les propriétés de tous les autres corpus, plus petits comme plus grands, sont toutes les mêmes, oral comme écrit.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

8 Discussion

On arrive donc à la conclusion qu'une partie importante de notre lexique (plus de la moitié de celui-ci) correspond à des termes que nous n'utilisons presque jamais. Globalement, ces items ne représentent pas non plus un grand volume de mots produits : de 1% à 5% selon les corpus (voir tableau 2). Néanmoins, cette faible fréquence signifie que nous produisons de 10 à 100 fois par heure des mots de ce type (selon le débit de production), mais que nombre de ces mots sont individuellement produits moins d'une dizaine de fois dans toute notre vie. Comment cela est-il possible ? Quelle maitrise avons-nous de ces items langagiers rares dont la présence est, globalement, massive dans la langue ?

La réponse à cette question ne peut être donnée dans le cadre de cet article. Cette recherche reste à mener. En particulier, il serait dangereux de spéculer ici sur certaines conséquences de la propriété des éléments rares sans vérifier in situ dans des corpus les plus denses possible les conditions d'usage des éléments rares.

Par exemple, il est possible qu'une proportion importante de ces éléments rares soit des items mal prononcés, mal entendus, mal répétés. Il est envisageable également que beaucoup de ces items soient des reprises ou répétitions en contexte d'items dont nous ne gardons jamais trace. On notera toutefois que l'on trouve les mêmes proportions dans le langage de la mère s'adressant à son enfant (voir le corpus de Thomas particulier parlant du fait de sa grande taille). Même s'il est fort probable que les parents et les locuteurs des enfants reproduisent de temps en temps les items erronés des enfants avec lesquels ils conversent, il est peu plausible que tous les items rares soient couverts de la sorte.

Il semble plus raisonnable de penser que, quelle que soit l'importance des micro-accidents du discours, une partie importante des items produits et entendus soient des items rares. Or la plupart des théories linguistiques s'attachent à décrire et comprendre les généralisations et les régularisations effectuées par les locuteurs. De ce fait, une large part des situations langagières reste en dehors du champ d'étude de la linguistique. On s'attache ainsi le plus souvent à décrire ou rechercher des universaux et non pas à identifier toutes les structures et situations particulières limitées à une personne et une occurrence. Même pour les travaux portant sur des analyses qualitatives fines, on sous-entend en général qu'il est possible d'extrapoler les analyses à d'autres situations.

Toutefois, si chaque événement langagier rare est unique ou presque, on peut imaginer qu'il existe dans le traitement des événements rares des principes généraux que l'on retrouve de manière systématique, mais qui portent sur la psycholinguistique plutôt que sur la linguistique. Au moins trois pistes théoriques et appliquées existent aujourd'hui pour étudier le domaine encore peu exploré des événements rares :

- la saillance ;
- l'acquisition de nouveaux mots ;
- la notion d'exemplaires.

Il est impossible de traiter de manière exhaustive (ni même superficielle) ces trois vastes domaines dans le cadre de la partie discussion d'un simple article, mais on peut les présenter de manière rapide et non exhaustive.

La notion de saillance est une notion largement utilisée en linguistique et dont la présence au sein de l'interaction de langage apparaît le plus souvent comme évidente. S'il existe de nombreux travaux sur la saillance, cette notion ne bénéficie pas d'une définition univoque (Geeraerts, 2000). Geeraerts explique ainsi qu'une unité linguistique n'est pas nécessairement plus ancrée parce que plus fréquente (et vice versa), car ce serait ignorer que les choix sémantiques des locuteurs sont indissociables de considérations pragmatiques et plus généralement contextuelles (2000, p. 75). Schmid (2007) propose une vision conceptuelle de la saillance : ce ne sont pas des entités réelles qui s'ancrent via des phénomènes de saillance, mais des entités appréhendées conceptuellement. Reste alors à définir et expliquer ce qu'est appréhender un concept. Quels que soient les arguments théoriques en jeu, on voit que l'on peut envisager que la saillance s'oppose à la fréquence dans le traitement des événements rares.

L'acquisition de nouveaux items par les enfants est un phénomène étudié depuis très longtemps (voir Clark, 1995, 2002, pour une revue complète du domaine par une de ses spécialistes). En particulier, le domaine de recherche a imaginé la notion de « *fast mapping* » pour décrire la capacité étonnante des enfants, après l'audition d'une seule occurrence d'un item nouveau, à repérer en un instant et se souvenir pendant au moins plusieurs semaines de la signification d'un item langagier. La notion de *fast mapping* (mise en correspondance rapide) est nécessaire pour expliquer la vitesse d'acquisition du lexique par les enfants. On pourrait aussi argumenter qu'elle s'applique en permanence, pour les enfants comme pour les adultes, pour nous permettre de traiter des événements rares.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

Enfin, la notion d'exemplaires, largement utilisée en phonologie (Pierrehumbert, 2002) ou en grammaire (Goldberg, 1995, 2006, Croft, 2001, Croft & Cruse, 2004, et beaucoup d'autres) se base sur le principe qu'un élément rencontré une fois va participer à la construction de l'ensemble des connaissances langagières. On va observer une généralisation par l'usage au fur et à mesure que les exemplaires sont rencontrés. Ce qui est intéressant dans cette théorie est que certains auteurs envisagent la possibilité de conservation en mémoire d'une quantité importante d'exemplaires. En particulier, on ne sait pas au moment où on entend un exemplaire s'il va être utilisé pour être généralisé plus tard. Il y a donc nécessairement une certaine conservation en mémoire des données langagières, quelle que soit la fréquence d'un item, en vu de sa réutilisation. Ce mécanisme pourrait être utilisé pour toute situation de langage et donc traiter les événements rares comme les événements fréquents.

9 Conclusion

Le langage de tous les jours, oral comme écrit, présente une répartition en fréquence des éléments utilisés qui suit une loi de Zipf. Une des conséquences de cette répartition est que nous passons une grande partie de notre temps à utiliser des éléments langagiers sur lesquels nous n'avons presque aucune connaissance et aucune expérience.

Cette caractéristique du langage peut être vue comme anecdotique ou fondamentale.

Anecdotique, elle ne représente que des situations qui n'ont pas d'influence sur les comportements usuels et sur l'intercompréhension. Nous disposons d'un ensemble de connaissances généralisées à partir de notre expérience ou faisant partie de notre bagage génétique : elles nous permettent de traiter toutes les situations exceptionnelles ou usuelles.

Fondamentale, elle est utilisée en permanence pour acquérir de nouvelles connaissances et entretenir les anciennes. C'est par là que passe l'apprentissage de mots nouveaux, mais aussi de toute connaissance langagière. C'est aussi par là que passe la création d'items nouveaux. Si ces items sont efficaces dans l'interaction, ils seront réutilisés avec succès. Mais quelque soit leur devenir, ils doivent passer par cette étape d'une première utilisation unique. Dans ces conditions, notre capacité à gérer les items rares et uniques serait fondamentale au même titre que notre capacité à utiliser efficacement les items fréquents : les deux capacités seraient en fait complémentaires dans le système du langage.

Quelles que soit la conclusion qui prévaudra dans quelques années, il apparaît que les événements rares forment un phénomène incontournable qui mérite de devenir un sujet de recherche fondamental. Espérons que ce sera le cas dans un avenir proche.

10 Bibliographie

Clark, E. V. (1995). *The Lexicon in Acquisition*. Cambridge University Press.

Clark, E. V. (2002). *First Language Acquisition*. Cambridge University Press.

Croft, W. (2001). *Radical construction grammar*. Oxford: Oxford University Press.

Croft, W., & Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge University Press.

Demuth, K., & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language*, *35*(01), 99–127. doi:10.1017/S0305000907008276

Geeraerts, D. (2000). Salience phenomena in the lexicon. A typology. In L. Albertazzi (Ed.), *Meaning and Cognition* (pp. 125-136). Amsterdam, Philadelphia: John Benjamins.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Mandelbrot, Benoit (1962). On the theory of word frequencies and on related Marko- vian models of discourse. In R. Jakobson (ed.), *Structure of Language and its Math- ematical Aspects*, pages 190–219. American Mathematical Society, Providence, RI.

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

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. doi:10.1080/00107510500052444

Pierrehumbert J. (2002). Word-specific phonetics, Gussenhoven C., Warnern (Eds.), *Papers in Laboratory Phonology VII*, p. 101–140, Mouton de Gruyter, Berlin, Germany, 2002.

Schmid, H.-J. (2007). Entrenchment, salience, and basic levels. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of Cognitive Linguistics* (pp. 117-138). Oxford: Oxford University Press.

Lieven, E., Salomo, D. & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. Cognitive Linguistics,20, 3, 481-508.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. Journal of Child Language, 28, 127-152.

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: how much is enough? *Journal Of Child Language*, 31(1), 101–121.

Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. Cambridge, MA: Addison-Wesley.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

Tableau 2 : Pourcentage de couverture pour les mots apparaissant de une à cinq fois dans l'ensemble d'un corpus – Paramètre α Loi de Zipf-Mandelbrot

			Types	Occ.	1 fois		2 fois		3 fois		4 fois		5 fois		Zipf
corpus	Lang.	Particip.	x 1000	x 1000	%Typ	%Occ	%Typ	%Occ	%Typ	%Occ	%Typ	%Occ	%Typ	%Occ	α
PL (tous)	Fra	Enfant	7	303	37	1	50	2	58	2	63	3	67	3	0,537
PL (tous)	Fra	Adulte	18	1 271	41	1	55	1	62	1	67	2	70	2	0,508
PL (moy)	Fra	Enfant	3	37	38	2	54	4	63	6	68	7	73	9	0,657
PL (moy)	Fra	Adulte	6	158	41	2	55	3	63	3	68	4	72	5	0,575
Man (tous)	Angl	Enfant	6	528	30	0	43	1	50	1	56	1	60	1	0,431
Man (tous)	Angl	Adulte	12	1 456	36	0	49	1	56	1	61	1	65	1	0,452
Man (moy)	Angl	Enfant	2	44	29	1	43	2	52	3	58	4	63	5	0,553
Man (moy)	Angl	Adulte	3	121	34	1	48	2	56	3	62	3	66	4	0,507
Thomas	Angl	Enfant	11	508	43	1	56	2	64	2	68	2	71	3	0,517
Thomas	Angl	Adulte	21	2 007	40	0	54	1	62	1	66	1	69	1	0,485
Oral OANC	Angl	Adulte	41	4 288	41	0	56	1	64	1	69	1	72	1	0,555
Ecrit OANC	Angl	Adulte	461	19 394	59	1	71	2	77	2	81	3	84	3	0,625
Gutenberg	Angl	Adulte	3 664	280 104	57	1	73	1	78	1	82	2	83	2	0,659

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801339

Tableau 3 : Comparaison corpus réduits et corpus complets sur les pourcentages de types et coefficient de Zipf

			%Typ 1 fois		%Typ 2 fois		%Typ 3 fois		%Typ 4 fois		%Typ 5 fois		α Zipf	
corpus	Lang.	Particip.	Tout	1/10	Tout	1/10	Tout	1/10	Tout	1/10	Tout	1/10	Tout	1/10
PL (tous)	Fra	Enfant	37	43	50	61	58	70	63	76	67	79	0,537	0,735
PL (tous)	Fra	Adulte	41	46	55	62	62	71	67	76	70	80	0,508	0,707
PL (moy)	Fra	Enfant	38	45	54	62	63	70	68	76	73	79	0,658	0,722
PL (moy)	Fra	Adulte	41	45	55	61	63	69	68	75	72	78	0,575	0,694
Man (tous)	Angl	Enfant	30	38	43	56	50	65	56	72	60	76	0,431	0,715
Man (tous)	Angl	Adulte	36	(35)	49	(51)	56	(59)	61	65	65	70	0,452	0,642
Man (moy)	Angl	Enfant	29	37	43	54	52	63	58	70	63	74	0,553	0,712
Man (moy)	Angl	Adulte	34	39	48	55	56	64	62	70	66	74	0,507	0,659
Thomas	Angl	Enfant	43	(41)	56	(55)	64	(64)	68	(69)	71	(73)	0,517	0,624
Thomas	Angl	Adulte	40	(38)	54	(52)	62	(60)	66	(66)	69	(69)	0,485	0,542
Oral OANC	Angl	Adulte	41	45	56	61	64	69	69	74	72	77	0,555	0,647
Ecrit OANC	Angl	Adulte	59	55	71	(69)	77	(75)	81	(79)	83	(82)	0,626	0,664
Gutenberg	Angl	Adulte	57	(53)	73	68	78	74	82	78	84	80	0,659	0,605