This et that dans les domaines spécialisés du corpus ICE-GB : quelles caractéristiques distributionnelles ?

Thomas Gaillat

This et *that* dans les domaines spécialisés du corpus ICE-GB : quelles caractéristiques distributionnelles ?

Thomas Gaillat
Université Paris-Diderot
Université Rennes 1

MOTS CLÉS
Démonstratif, déterminant, domaine spécialisé, pro-forme, *that*, *this*.

RÉSUMÉ
Cet article analyse les caractéristiques distributionnelles des deux démonstratifs « this *» et «* that *» afin d'identifier des usages spécifiques en fonction de domaines spécialisés de l'anglais. Les données sont collectées dans le corpus ICE-GB. L'étude consiste à échantillonner le corpus en sous-corpus, en fonction du domaine spécialisé et du mode écrit ou oral des textes. Les sous-corpus relevant de l'anglais général sont distingués de ceux relevant de domaines spécialisés (médecine, science et technologie). Pour chaque sous-corpus, l'outil ICECUP est utilisé pour effectuer des requêtes et extraire le nombre d'occurrences des démonstratifs en fonction de leur catégorie grammaticale : déterminant, pro-forme, adverbe, complétif et pronom relatif. La distinction par catégorie vise à compter le nombre exact d'une forme particulière en fonction d'un sous-corpus spécifique représentant un domaine. Les résultats statistiques globaux montrent une corrélation limitée entre les démonstratifs et les domaines spécialisés. Cependant, certaines catégories grammaticales entretiennent un lien étroit avec un domaine particulier. Enfin, l'étude montre des tendances marquées concernant l'usage des démonstratifs dans leur rôle de pro-forme.*

KEY WORDS
Demonstrative, determiner, pro-form, specialized domain, *that, this*.

ABSTRACT
This paper analyzes the distributional characteristics of demonstratives "this" and "that" so as to identify specific uses in relation to specialized domains of English. Data were collected from the ICE-GB corpus. The study consisted in extracting – oral or written – samples from sub-corpora linked to a specialized domain. The author classified the sub-corpora according to whether they corresponded to general English or to the medical, scientific and technological domains. For each sub-corpus, ICECUP was used to make queries and extract the number of occurrences of the demonstratives according to their syntactic function. Grammatical function (i.e., determiner, pro-form, adverb, complementizer or relative pronoun) was used to establish exact counts of a particular form in relation to a specific sub-corpus representing a specialized domain. The statistical results show a limited correlation between the demonstratives and the domains at a global level, but some specific functions do have a substantial link with particular

specialized domains. The study reveals marked trends regarding the use of demonstratives in their pro-form function.

1. Introduction

L'acte d'enseignement en anglais de spécialité soulève des questions dont certaines solutions trouvent leur source en didactique et en linguistique. Pour construire des dispositifs d'enseignement LANSAD, il est nécessaire de s'appuyer sur des connaissances linguistiques et culturelles propres au domaine de spécialité. Ces connaissances sont constituées à la fois d'éléments pédagogiques spécifiques tels que les conditions d'apprentissage et le type de public, mais aussi d'éléments linguistiques spécifiques des domaines spécialisés auxquels la spécialité enseignée se rattache. Nous nous rallions à Michel Petit (2010 : §20) qui définit chaque domaine spécialisé comme

> un secteur de la société constitué autour et en vue de l'exercice d'une activité principale qui, par sa nature, sa finalité et ses modalités particulières ainsi que par les compétences particulières qu'elle met en jeu chez ses acteurs, définit la place reconnaissable de ce secteur au sein de la société et d'un ensemble de ses autres secteurs et détermine sa composition et son organisation spécifiques.

Comprendre les éléments constitutifs du discours d'un domaine spécialisé constitue un enjeu important. Cette compréhension permet ainsi de nourrir, à la manière des études portant sur des variations linguistiques en fonction des genres discursifs (Hyland 2010), d'éventuelles applications didactiques dans l'enseignement de l'anglais de spécialité (Parkinson & Adendorff 2004).

Notre expérience auprès d'étudiants LANSAD des domaines médical, scientifique et technologique nous a permis de constater la difficulté qu'éprouvent les apprenants à construire des textes cohésifs. Nous nous référons ici à Halliday et Hasan (1976), selon qui la cohésion d'un texte dépend, entre autres, du concept de référence, deux de ses constituants majeurs étant *this* et *that*. Une étude menée sur le corpus d'apprenants LONGDALE (Gaillat 2013) aborde la problématique de la construction de la référence chez les apprenants. Elle met en évidence des difficultés de choix entre les deux formes, mais aussi entre l'une ou l'autre des formes et le pronom personnel *it* ou l'article *the*. Afin de mieux appréhender l'usage fait par les apprenants des démonstratifs pour construire leur discours en langue de spécialité, il paraît nécessaire d'analyser, dans un premier temps, leur usage par les locuteurs natifs en anglais général et en fonction de certains domaines spécialisés (médecine, science et technologie). Fondée sur des données chiffrées, cette analyse fait apparaître des tendances d'utilisation de *this* et de *that* selon le domaine spécialisé. Les observations réalisées dans le cadre de cette étude auront peut-être à l'avenir des applications didactiques.

La recherche entreprise vise à dresser la carte des réalisations de *this* et de *that* en fonction de domaines spécialisés scientifiques. Pour ce faire, trois catégories de textes écrits et oraux ont été extraits du corpus *International Corpus of English – Great Britain* (ICE-GB) : les textes d'anglais général, les textes du domaine médical et ceux du domaine scientifique et technologique. Pour la suite de l'article, et afin de favoriser la clarté de notre propos, nous utilisons le terme de sous-corpus pour désigner ces catégories, bien que cet usage soit quelque peu imprécis à propos de l'anglais général, car il s'agit du corpus entier. L'objet de cette recherche est d'identifier certaines tendances dans l'usage des démonstratifs, selon qu'ils sont utilisés dans un domaine spécialisé ou dans un autre. Nous focalisons notre propos sur les usages en pro-forme et déterminant, codés distinctement dans le corpus ICE-GB (Nelson, Wallis & Aarts 1998). Cependant, nous élargissons notre analyse des données aux autres catégories grammaticales auxquelles peuvent appartenir les formes, à savoir les usages complétif et relatif de *that* d'une part, et l'usage adverbial des deux formes d'autre part.

Les travaux antérieurs abordent *this* et *that* selon deux perspectives. La première concerne les réalisations fonctionnelles des deux formes. Les positions syntaxiques qu'elles peuvent occuper sont doubles. On peut retrouver *this/that* devant un syntagme nominal ou en position de syntagme nominal. Les Anglophones utilisent les termes *modifier* pour la position ante-nominale et *head* pour la position nominale. Quirk, Leech et Svartvik (1985) distinguent usage en déterminant et usage nominal ; Fraser et Joly (1979) distinguent les pronoms supplétifs et les pronoms complétifs. Pour notre part, nous reprendrons les termes de déterminant et de « pro-forme » (Lapaire & Rotgé 2008 : 50-51). La catégorie de « déterminant » permet d'affirmer la position ante-nominale des formes lorsqu'elles participent au processus de détermination d'un nom. « Pro-forme » permet d'indiquer les formes *this/that* qui apparaissent sans nom et qui jouent le rôle de syntagme nominal. En outre, la dénomination « pro-forme » éclaire sur le rôle de référence sémantique d'une forme qui peut être plus étendue qu'un groupe nominal isolé.

La seconde perspective concerne la fonction référentielle, plus particulièrement le domaine de la deixis. Traditionnellement, les deux formes permettent d'exprimer la distance qui sépare le locuteur de l'objet auquel il/elle fait référence. Cette valeur est reprise par Biber (1999 : 347) selon qui la distinction s'opère en fonction de la distance du référent. Une analyse plus approfondie permet de situer les deux formes en fonction du type de référence qu'elles mettent en œuvre. La distinction endophore/exophore est développée par Halliday et Hasan (1976) et reprise par Fraser et Joly (1979 : 105) qui ajoutent la notion de sphère du locuteur comme repère. *This* place le référent dans la sphère du locuteur quand *that* le place en dehors. Lapaire et Rotgé (2008 : 59) reprennent l'idée selon laquelle le « moi »

énonciateur indique que le référent, désigné par le démonstratif, s'approche ou s'éloigne. Ils proposent l'idée de clôture avec *that* et de non-clôture avec *this*. *This* traduit le « refus ou impossibilité de traiter comme clos l'acquis opérationnel marqué par TH- ». *That* dénote « l'acceptation, bouclage, ou, plus généralement, clôture de l'acquis opérationnel marqué par TH » (Lapaire & Rotgé 2008 : 64). Le rôle de l'énonciateur est central dans leur explication. Huddleston et Pullum (2002 : 1504-1510)[1] s'inscrivent, eux aussi, dans l'architecture exophore/endophore en attribuant des valeurs anaphorique ou déictique aux deux formes. L'anaphore reprend ce qui se trouve dans le co-texte, quand le déictique renvoie à des objets présents dans la situation d'énonciation. Ce type de reprise situationnelle tend à favoriser la valeur de proximité ou d'éloignement dans le processus de référence et à ne pas prendre en compte l'énonciateur comme responsable des opérations cognitives construisant la référence, ce qui est central chez les auteurs francophones mentionnés (Lapaire & Rotgé ; Fraser & Joly ; Cotte). Kleiber (1992), quant à lui, différencie le déictique de l'anaphorique, selon qu'il s'agit d'information nouvelle ou non (et non plus en fonction de la localisation situationnelle ou contextuelle). Cotte (1993 : 47) réconcilie les deux points de vue en montrant que ce qui est déjà présent pour le locuteur peut être perçu comme « nouveau dans le rapport interlocutif » et, donc, nouveau pour l'allocutaire : il y a une valeur de reprise anaphorique si l'on se place du point de vue du locuteur et une valeur déictique si l'on se place du point de vue de l'allocutaire. Bordet (2011) relève cette ambivalence dans son travail sur *this* lorsqu'elle indique sa capacité à faire progresser la focalisation tout en assurant l'établissement de la référence. Ce faisant, elle rejoint ainsi Cornish (1999 : 31-32) qui considère les procédures déictiques et anaphoriques comme des fonctions du discours « constituant un continuum sur le plan de l'indexicalité ». Dans la suite de cet article, nous nous référons aux deux formes en utilisant le terme historique, et peut-être plus neutre, de « démonstratifs ». Nous réservons les termes « anaphorique » et « déictique » pour les explications relatives au fonctionnement référentiel des démonstratifs.

Après la catégorisation du corpus ICE-GB, des relevés ont été effectués pour déterminer les fréquences d'occurrence de *this* et de *that*, dans le cadre de notre projet, qui consiste à identifier les corrélations qui existent entre les domaines spécialisés de discours et ces marqueurs. La première partie de l'article est consacrée aux données, à leur collecte et à leur normalisation. Dans un deuxième temps, nous menons une analyse statistique sur les profils de corpus en fonction des différentes catégories grammaticales de *this* et de *that.* Nous tentons d'identifier le rôle que peut jouer un *this* ou *that* en fonction d'une catégorie donnée et des domaines spécialisés. L'évaluation statistique de la distribution de *this* et *that* est ensuite

[1] Le chapitre a été écrit par Lesley Stirling et Rodney Huddleston.

mise en œuvre. La troisième partie de l'article propose un retour sur les données et analyse la distribution textuelle de *this* et de *that* au sein de textes médicaux écrits et oraux. Elle permet de vérifier les observations effectuées à l'issue de l'analyse statistique et de définir les bases de ce qui paraît être la prochaine étape de la linguistique textuelle : caractériser les étapes obligées des structures textuelles par leurs marqueurs de prédilection. De ce point de vue, l'anglais de spécialité, en particulier dans des écrits extrêmement codés tels que les résumés ou les articles scientifiques, constitue un observatoire privilégié.

2. Méthodologie

Il s'agit ici de décrire le corpus ICE-GB et la structuration des données le constituant. Sont présentés ensuite la manière dont sont extraits et catégorisés en sous-corpus les échantillons comprenant *this* et *that.*

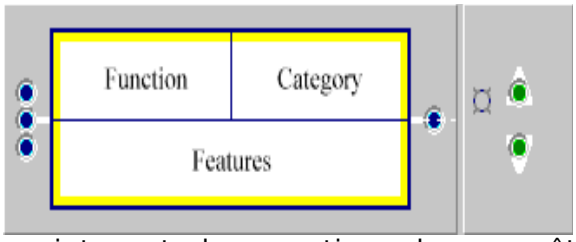
2.1. Description du corpus ICE-GB

Le corpus ICE-GB est composé de 1 061 263 mots. Aarts, Nelson et Wallis (2007) précisent qu'il comprend 500 textes d'environ 2 000 mots chacun, qui proviennent de la langue orale ou écrite et ont été collectés entre 1990 et 1996 : leurs auteurs étaient alors âgés de 18 ans et plus. Ils ont été catégorisés de manière à offrir une traçabilité du type de texte. Par exemple, un certain nombre d'entre eux, utilisés dans la présente étude, sont classés selon le chemin suivant : *Spoken > Dialogues > Public > Classroom lessons*. Chaque texte possède une fiche caractéristique précise sur son auteur, son lieu et date de collecte et sa thématique de contenu. Un large éventail de catégories caractérise ces textes, qu'il s'agisse du mode oral ou écrit. À l'oral, on trouve des dialogues enregistrés lors de conversations directes ou téléphoniques, ou lors d'émissions de radio, ou encore lors de séances parlementaires. Le corpus oral comprend aussi, entre autres, des monologues enregistrés lors de conférences ou de leçons en classe. Le mode écrit comporte des textes universitaires et scientifiques dans les domaines des sciences humaines, des sciences sociales et des technologies. Le corpus comprend des écrits journalistiques, administratifs, ou encore littéraires (extraits de romans). Afin d'explorer le corpus, le logiciel ICECUP[2] a été développé et propose plusieurs outils, dont un outil de navigation appelé *Corpus map*. Il permet d'évoluer dans les catégories et sous-catégories à la manière d'un explorateur de fichiers (Nelson, Wallis & Aarts 2002 : 87). En parcourant l'ensemble des textes classés dans la catégorie *Classroom lessons*, par exemple, quatre leçons au contenu médical totalisant 8 249 mots sont identifiées.

[2] <http://www.ucl.ac.uk/english-usage/resources/icecup/index.htm>. Consulté le 6 mai 2013.

Outre une classification thématique des textes, le corpus ICE-GB propose une annotation syntaxique de ceux-ci, sous forme d'arbres permettant d'effectuer des requêtes à l'aide d'un autre outil du logiciel ICECUP. Avant d'aborder les requêtes, il est important d'expliquer la structuration de l'annotation syntaxique et des parties du discours élaborée par les auteurs du corpus ICE-GB. La figure 1 (Nelson, Wallis & Aarts 1998) décrit la manière dont se structure l'annotation « partie-du-discours » sous forme de « nœuds ». Une nomenclature précise est utilisée pour décrire chaque unité[3]. Un premier élément identifie la fonction (par exemple, *Phrase Unit, Adverb Head, Noun Phrase Head, Subordinate Head*), c'est-à-dire le rôle que joue chaque mot par rapport à son contexte immédiat. Sa catégorie grammaticale est ensuite spécifiée. Par exemple, *this* et *that* pourront se voir assigner les étiquettes PRON pour *pronoun* ou DTCE pour *Central Determiner*. Le troisième élément concerne des caractéristiques particulières que peuvent prendre les formes. Concernant *this* et *that,* il s'agira par exemple de l'étiquette DEM pour démonstratifs. Chacun de ces nœuds s'articule avec les autres, de telle manière qu'une représentation syntaxique par arborescence est possible. Cette arborescence résulte d'une analyse automatique faite par un *parser*, développé à l'Université de Nijmegen, qui utilise une grammaire formelle pour analyser chaque unité au niveau des mots, des groupes de mots et des propositions (Aarts, Nelson & Wallis 2007 : 29).

Figure 1. Annotation « partie-du-discours » du corpus ICE-GB



Abordons maintenant la question des requêtes utiles pour l'extraction des occurrences de *this* et de *that*. La structuration des données, sous forme d'annotation syntaxique, sert de point de départ pour élaborer des requêtes sur le corpus. Il s'agit de convertir les catégories grammaticales, sur lesquelles l'utilisateur souhaite faire une recherche, en requêtes logiques compatibles avec la structuration des données du corpus ICE-GB. Les *Fuzzy Tree Fragments* ou FTF, outils de recherche par motif du logiciel ICECUP, permettent de modéliser chaque requête de façon intuitive. Leur principe de fonctionnement repose sur la structuration des nœuds et

[3] L'accès aux fonctions, catégories et traits (*Features*) se fait à l'adresse suivante <http://www.ucl.ac.uk/english-usage/resources/grammar/index.htm>. Consulté le 6 mai 2013.

sur leurs articulations. Plutôt que d'utiliser le langage *Logic*, d'un abord complexe, l'utilisateur construit un motif du type du fragment d'arbre syntaxique qu'il souhaite trouver. Les liens plus ou moins stricts de domination et d'adjacence entre les nœuds sont définis et les catégories, fonctions et caractéristiques de chaque nœud sont renseignées partiellement ou de manière exhaustive. Des mots peuvent être ajoutés au motif FTF qui traduit la requête en langage *Logic*. Après validation, toutes les occurrences dont les mots et l'annotation correspondent au motif seront affichées avec leur contexte proche. La section 2.2 aborde la syntaxe exacte des requêtes élaborées pour les besoins de l'étude.

2.2. Extraction et catégorisation en sous-corpus

À présent que le corpus ICE-GB, sa structuration et ses outils ont été exposés, il s'agit de décrire la première étape de cette étude. Elle consiste à subdiviser le corpus en trois sous-corpus sur lesquels seront effectuées des requêtes d'ordre syntaxique qui permettront d'extraire les occurrences de formes en fonction de schémas syntaxiques. Le premier niveau de division concerne les modalités orales ou écrites. Pour chacun de ces deux ensembles, nous distinguons trois types de séries de textes. La première concerne les textes dont les contenus sont médicaux. La deuxième série concerne les textes scientifiques et technologiques (textes médicaux exclus) et la troisième série correspond à tous les textes du corpus ICE-GB (y compris ceux des deux premières séries), et s'inscrit donc dans le domaine de l'anglais général. Chaque série est identifiée à partir des données sociolinguistiques (discipline de l'auteur, institution) fournies pour chacun des textes. Il y a au final trois sous-corpus pour chaque mode oral ou écrit, soit un total de six sous-corpus.

L'outil d'exploration ne permet pas de procéder à des regroupements *ad hoc* qui autoriseraient une seule manipulation pour la sélection de l'ensemble des textes correspondant à un contenu particulier. L'outil *Corpus map* permet d'identifier l'ensemble des textes scientifiques du corpus et leur nombre de mots. Chaque texte est trié manuellement en fonction du domaine scientifique dont il relève. Ce processus permet, par addition, d'obtenir le nombre d'occurrences de mots par catégorie (tableau 1). Concernant l'origine des quatre sous-corpus médicaux, scientifiques et technologiques, ceux-ci comprennent des textes classés dans les catégories suivantes : leçons (*classroom lessons*), monologue, démonstrations scientifiques non rédigées (*unscripted demonstrations*) et monologue, discours non rédigés (*unscripted speeches*) pour le mode oral ; écrits scientifiques et universitaires (*academic writing*) et écrits non scientifiques (*non-academic writing*), chaque catégorie du mode écrit incluant les sous-catégories « sciences naturelles » et « technologie » (*natural sciences and technology*).

Tableau 1. Ensemble des sous-corpus extraits du corpus ICE-GB

Discourse	corpus subset size
Oral medical	15178
Oral science sci & tech (not med)	20680
Oral general	637682
Written medical	8431
Written sci & tech (not med)	38202
Written general	423581

Une fois ces catégories établies, la deuxième étape de l'étude consiste à élaborer des requêtes pour faire le décompte du nombre d'occurrences de chaque démonstratif singulier et pluriel en fonction de sa catégorie grammaticale. Il s'agit de formaliser de manière logique des requêtes grâce à l'outil du logiciel ICECUP (Nelson, Wallis & Aarts 2002). Ce travail se situe dans le prolongement d'une étape préliminaire d'identification des catégories grammaticales auxquelles peuvent correspondre ces deux formes. Les travaux de recherche sur le sujet, présentés dans l'introduction de cet article, permettent d'avancer cinq catégories grammaticales possibles pour *that* et trois pour *this.* Elles sont résumées dans le tableau 2, avec des exemples tirés du corpus ICE-GB. Bien que de nombreux auteurs aient glosé sur la dimension sémantique prise par les deux formes, notamment en ce qui concerne le domaine de la deixis (cf. introduction), cette étude se concentre sur leur rôle syntaxique, en relation avec l'annotation du corpus.

Tableau 2. Les différentes catégories grammaticales de *this* et de *that*

Forme	**Catégorie grammaticale**	**Définition**	**Exemple**
This	pro-forme	Remplace une forme telle que prédicat, proposition, phrase, conglomérat d'énoncés, nom	*This* is not like a real situation (SIB-004-136)
This	adverbial	Modifie un adjectif	When you felt *this* strong […] (W2F-008-038)
This	déterminant	Introduit un syntagme nominal	From the actual target tissue which in *this* particular case we are talking about is muscle [...] (S1B-009-192)
That	pro-forme	Remplace une forme telle que prédicat, proposition, phrase, conglomérat d'énoncés, nom	*That* sounds reasonable (S1B-003-019)
That	déterminant	Introduit un syntagme nominal	and yet *that* language was being channelled into a very narrow, field (S1A-001)
That	adverbial	Modifie un adjectif	And it's not *that* far anyway (S1A-006-294)
That	complétif	Introduit une subordonnée, complément d'un verbe	It means *that* the zero level is variable so that you get readings […] (S1B-004-252)
That	relatif	Introduit une subordonnée complétant un antécédent	But it's random variation as we said before *that* will tend to even out [...] (S1B-004-346)

Les requêtes sont élaborées à l'aide de l'outil *Logic* de ICECUP. Par exemple, afin d'identifier les occurrences de *this* pro-forme, nous décrivons le motif correspondant à toutes les occurrences du mot *this*

en position de tête de syntagme nominal (on ne renseigne que la zone *Function* avec NPHD pour *Noun Phrase Head*). Traduite en langage *Logic* (Aarts, Nelson & Wallis 2007 : 30), la requête prend la forme *this*+<NPHD> et suffit à isoler tous les *this* pro-formes ayant pour catégorie « pronom » (pron) et pour caractéristique « démonstratif » (dem). Plus simplement, la requête *this*+<DTCE> donne l'ensemble des formes de *this* déterminant. Concernant *that,* et du fait de ses cinq catégories grammaticales possibles, certaines requêtes sont plus détaillées, notamment celles correspondant au *that* pro-forme. La requête *that*+<NPHD> couvre les formes « démonstratif » et « pronom relatif ». S'il s'agit d'isoler l'une de l'autre, les deux requêtes suivantes doivent être saisies : *that*+<NPHD, PRON(dem)> et *that*+<NPHD, PRON(rel)>. C'est donc grâce à l'utilitaire de recherche du logiciel, associé à un langage de requête simple, qu'il est possible d'extraire l'ensemble des formes et d'obtenir des fenêtres de résultats les présentant en contexte.

La collecte des occurrences de *this* et de *that* peut donc s'effectuer texte par texte, et permet d'obtenir un tableau synthétique des fréquences absolues par sous-corpus. Du fait de l'irrégularité de taille de chaque sous-corpus, il convient, avant toute analyse, de normaliser ces résultats en suivant Gries (2010). Afin de prendre en compte la différenciation oral/écrit par domaine spécialisé, essentielle à l'analyse, il est nécessaire de normaliser les fréquences au regard de chacun des sous-corpus oral et écrit du corpus ICE-GB. Par exemple, le sous-corpus oral médical comprend 15 178 mots et le sous-corpus écrit en totalise 8 431. Cette normalisation permet de produire les fréquences relatives par million de mots pour chaque occurrence et pour chaque catégorie grammaticale, à la manière des deux exemples suivants :

This pro-forme - médical oral : $\frac{47 * 1\,000\,000}{15178} = 3\,097$

This pro-forme - médical écrit : $\frac{36 * 1\,000\,000}{8431} = 4\,151$

Une fois effectué ce travail de normalisation, le travail d'analyse des données commence pour déterminer les différences entre les corpus et, donc, entre les domaines spécialisés et l'anglais général, mais aussi entre les catégories grammaticales des deux formes.

3. Analyse et résultats

Les données collectées et normalisées sont consignées dans un tableau synthétique manipulable et exploitable permettant de répondre à plusieurs questions. La première consiste à s'interroger sur les différences qui existent entre les deux formes sans tenir compte de leur catégorie grammaticale. La deuxième porte sur la mise en regard des sous-corpus en fonction des formes et de leur

catégorie grammaticale, en prenant également en compte les différences comparatives. Une dernière question concerne la validité des observations par rapport à la loi du χ^2. Les variations des données analysées suivent-elles cette loi ou sont-elles dues à des facteurs propres aux catégories choisies ? Le test du χ^2 permet de signaler une corrélation entre les catégories de données.

3.1. Approche globale

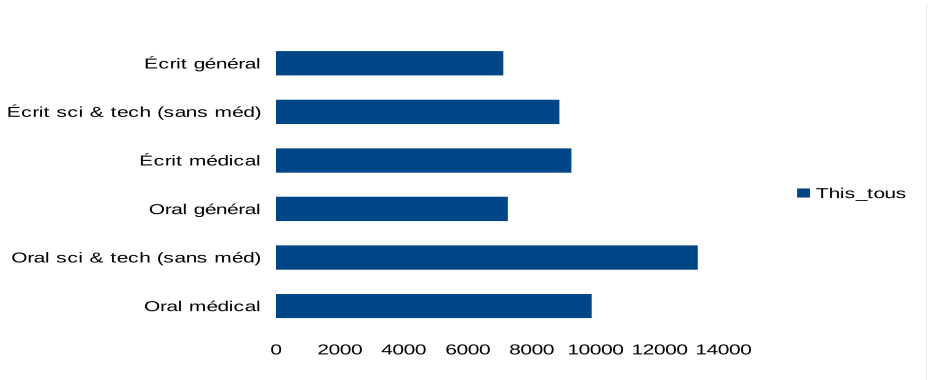
L'approche globale consiste à prendre en compte l'ensemble des formes de *this* et de *that* et à observer leur distribution sans distinction, c'est-à-dire sans tenir compte de leurs catégories grammaticales particulières. Il s'agit des parties grisées du tableau 3. Celui-ci regroupe l'ensemble des données relatives aux usages de *this* et de *that* selon leur catégorie grammaticale et selon le sous-corpus.

Tableau 3. Données normalisées : fréquences relatives observées dans les parties orale et écrite du corpus ICE-GB par million de mots

Discourse	This_all	this proform	this determiner	this adverbial	that_all	that proform	that determiner	that adverbial	that complementizer	that relative pron
Oral medical	9883	3097	6786	0	25827	8301	4151	2570	5996	4810
Oral science sci & tech (not med)	13201	4932	8269	0	23743	6867	3385	2466	5319	5706
Oral general	7259	2744	4509	6	23803	7761	2895	2275	6768	4104
Written medical	9252	4151	5100	0	12573	2728	830	0	5337	3677
Written sci & tech (not med)	8874	3586	5288	0	8638	1361	1021	0	4607	1649
Written general	7120	2828	4287	5	10699	1598	958	19	6209	1915

Pour un premier niveau d'analyse, le calcul des fréquences relatives totales de *this* et de *that* observées dans les parties orale et écrite du corpus ICE-GB est effectué (cf. figure 2).

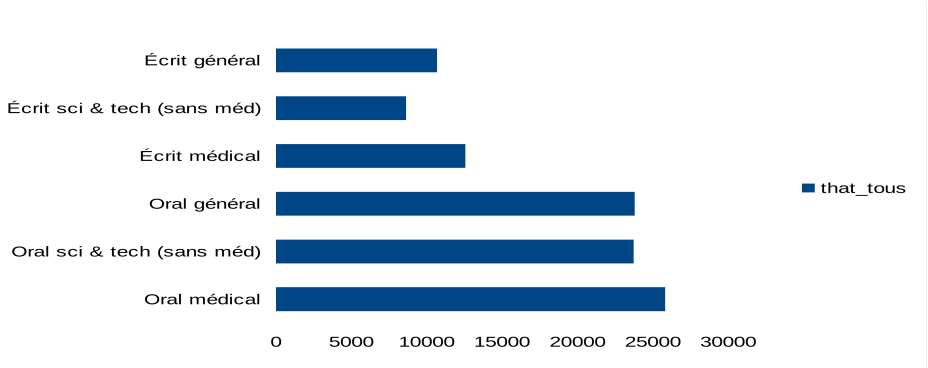
Figure 2. Fréquences relatives de *this* dans le corpus ICE-GB



On observe clairement une équivalence des usages au sein de chaque mode oral ou écrit, excepté dans les sous-corpus scientifiques et technologiques, qui présentent des fréquences d'utilisation plus élevées. Celles-ci peuvent s'expliquer par une hétérogénéité thématique relative à l'intérieur de ces deux sous-corpus, l'ensemble oral comportant peu de textes technologiques, et une prépondérance

de textes proches des sciences naturelles. À l'inverse, le sous-corpus écrit comprend une part importante de textes technologiques. Par conséquent, il est possible que les textes technologiques comportent une variation que l'on retrouve dans cet histogramme. Si l'on compare simplement les sous-corpus médical et général, on trouve un usage équivalent de *this* entre oral et écrit.

Illustration 3. Fréquences relatives de *that* dans le corpus ICE-GB



Les observations de *that* en fonction des différents sous-corpus (cf. figure 3) révèlent des différences notables entre chaque mode oral ou écrit. *That* est systématiquement sous-utilisé à l'écrit par rapport à l'oral. Il y a donc deux tendances inverses entre *this* et *that.* Si l'on considère les différentes catégories grammaticales des deux formes, on voit que *that* recouvre les mêmes catégories que *this* (pro-forme, déterminant, adverbial) mais en couvre d'autres très importantes à l'oral, comme la forme complétive que l'on retrouve dans les énoncés de type *I think that...* et la forme pronom relatif. En termes schématiques, *this* permet seulement la reprise d'éléments du discours, quand *that* va au-delà, en permettant la complexification du discours grâce à des structures syntaxiques fondées sur l'hypotaxe, prépondérante à l'écrit.

3.2. Approche détaillée

En reprenant les résultats détaillés du tableau 3 et en les comparant avec les résultats obtenus par Biber (Biber *et alii.* 1999 : 274) et publiés dans la LGSWE[4], on peut faire les observations suivantes concernant l'usage en déterminant des formes (cf. tableau 4). On prendra soin de noter que les relevés de Biber concernent des conversations quand ceux du corpus ICE-GB comprennent aussi des monologues. Là où Biber *et alii.* relèvent une fréquence de 2 500 *this* par million de mots à l'oral (singulier et pluriel) en conversation, le relevé dans le corpus ICE-GB en comprend 4 500 dans le domaine de l'anglais général à l'oral. En ce qui concerne les occurrences de *that* (singulier et pluriel) en fonctionnement déterminant, le corpus ICE-GB donne une fréquence

[4] *Longman Grammar of Spoken and Written English.*

de 2 900 contre 2 500 dans l'étude de Biber *et alii*. Le comparatif des usages en pro-forme donne les observations suivantes. Biber *et alii.* (1999 : 349) rapportent une fréquence de 2 000 *this* (singulier et pluriel) en conversation. Le relevé dans le corpus ICE-GB en donne 2 700 dans le domaine de l'anglais général à l'oral. *That* (singulier et pluriel) en conversation totalise une fréquence de 11 500 dans l'étude de Biber *et alii.*, quand le relevé du corpus ICE-GB en montre 7 800 dans le domaine de l'anglais général à l'oral. En résumé, le comparatif des pro-formes montre plus de *that* dans l'étude de Biber *et alii..* que dans le corpus ICE-GB, et moins de *this* que dans le corpus ICE-GB. Au niveau des déterminants, le corpus ICE-GB donne plus de *this* que la LGSWE et une fréquence similaire entre les deux pour *that*.

Tableau 4. Comparatif des formes entre le corpus de Biber et le corpus ICE-GB

	Biber	ICE-GB
This (déterminant)	2 500	4 500
That (déterminant)	2 500	2 900
This (pro-forme)	2 000	2 700
That (pro-forme)	11 500	7 800

Le domaine de l'anglais général à l'oral général du corpus ICE-GB inclut non seulement des conversations, mais aussi des monologues et leçons, ce qui peut expliquer cette différence. En effet, le monologue, auquel s'apparente aussi la leçon, a souvent pour objectif l'explication d'un concept en situation, ce qui peut impliquer un rapprochement dudit concept vers la sphère de l'énonciateur de manière physique ou mémorielle. Ceci est illustré dans l'exemple identifié S2A-051 : *You 'll see creatures like* this *one and if you try and capture it you can spend ages and ages watching them*. Il s'agit d'une leçon de biologie, et le locuteur focalise son propos sur le terme *creatures*, élément appartenant à son environnement, ce qui implique l'utilisation de *this*. Le monologue correspond lui aussi à une construction de discours dont la cohésion dépend de l'environnement textuel et physique organisé par l'auteur du discours. Il en découle un grand nombre de référents étroitement liés au locuteur de manière déictique ou anaphorique, d'où l'usage de *this*.

Inversement, le fait que *that* soit moins fréquent dans le corpus ICE-GB que dans la LGSWE peut aussi provenir de la part importante des monologues et leçons dans le corpus. La conversation implique un co-locuteur. Contrairement au monologue, où le locuteur fait référence à des concepts qu'il fait siens, la conversation implique des reprises de concepts portés par le co-locuteur. Ces reprises signalent probablement, le plus souvent, une acceptation sans modification des caractéristiques du référent, ce que *that* est apte à faire. L'exemple

suivant, extrait (S1A-012) des conversations directes du corpus, illustre le rôle joué par *that* dans un dialogue.

> Locuteur C : « I 'm not really in favour of boys' schools taking on girls in the sixth form and *that* 's it. It would be a very odd compromise »
>
> Locuteur D : « Uhm yes »
>
> Locuteur E : « Uh uhm »
>
> Locuteur C : « Added to which of course it destroys girls' schools »
>
> Locuteur B : « Yes it certainly does *that* »

La première occurrence de *that* permet au locuteur d'affirmer son point de vue et de clore le débat et la seconde permet la reprise par le locuteur B de l'argument avancé par le locuteur C. Selon Lapaire et Rotgé, il y a « acceptation, bouclage, ou, plus généralement, clôture (-AT) de l'acquis opérationnel marqué par TH » (2008 : 64) avec *that*. Le fait que le corpus ICE-GB contienne une grande part de monologues tend à réduire la part relative des conversations au regard du corpus de Biber *et alii*. Ceci peut expliquer la relative infériorité numérique des occurrences de *that* dans le corpus ICE-GB

Considérant les deux tendances de *this* et de *that*, et afin d'affiner les rôles que les deux formes jouent, il est nécessaire de procéder à une analyse plus détaillée des sous-corpus en fonction des catégories grammaticales. Le calcul des profils par colonne et par ligne du tableau 3 donne les résultats suivants.

Tableau 5. Profils par colonne : importance relative des sous-corpus pour chaque catégorie grammaticale (pourcentage)

Discourse	this proform	this determiner	this adverbial	that proform	that determiner	that adverbial	that complementizer	that relative pron
Oral medical	15	20	0	29	31	35	18	22
Oral science sci & tech (not med)	23	24	0	24	26	34	16	26
Oral general	13	13	57	27	22	31	20	19
Written medical	19	15	0	10	6	0	16	17
Written sci & tech (not med)	17	15	0	5	8	0	13	8
Written general	13	13	43	6	7	0	18	9

Les profils par colonne (tableau 5) permettent de comparer les sous-corpus entre eux, pour une forme donnée, et de voir si certaines catégories grammaticales sont plus fréquentes dans un sous-corpus que dans un autre. En d'autres termes, on mesure l'attractivité qu'exerce un sous-corpus par rapport aux autres, pour une occurrence de *this* ou de *that* d'une catégorie grammaticale donnée. Par exemple, on peut noter que le l'emploi de *that* qui correspond aux fonctions pro-forme et déterminant est plus fréquent dans les trois sous-corpus oraux que dans les sous-corpus écrits. Les sous-corpus oraux sont donc plus attractifs vis-à-vis des *that* pro-forme et déterminant que les sous-corpus écrits. Cette observation corrobore celle de Biber *et alii.* qui notent que la distribution des déterminants

et des pronoms démonstratifs est plus importante dans le registre de la conversation qu'à l'écrit (Biber *et alii.* 1999 : 274 ; 349). Ceci peut s'expliquer par le fait que *that* est un outil anaphorique qui permet au locuteur la reprise de concepts déjà mentionnés (dans son discours). Le mode de l'oral impose une charge cognitive plus importante quant à la mémorisation des différents référents du discours. Par conséquent, l'utilisation de *that* permet une réactivation des référents en fonction des énoncés dans lesquels ils apparaissent.

Le cas de *this* répond à deux variations. En fonctionnement déterminant, *this* est plus présent à l'oral et, notamment, dans les sous-corpus scientifique, technologique et médical. En fonctionnement pro-forme, *this* est autant présent à l'oral qu'à l'écrit, notamment dans le sous-corpus scientifique et technologique et le sous-corpus médical. La langue scientifique a pour objectif l'observation, la démonstration et la leçon. Ce type de déroulement discursif favorise donc la mise au premier plan de concepts déjà exprimés dans le discours ou déjà connus. Le locuteur focalise son propos sur une entité pour développer son raisonnement en ajoutant de nouvelles informations sur le concept posé en amont. Shirley Carter-Thomas et Elisabeth Rowley-Jolivet (2001) soulignent que la dimension déictique de *this* sert à pointer vers un nouvel élément dans les présentations scientifiques à l'oral. Le référent porte de nouvelles informations et se trouve dans la sphère du locuteur (Fraser & Joly 1979) d'où l'utilisation de *this.* Ce référent étant en cours d'analyse, il y a « impossibilité de traiter comme clos l'acquis opérationnel marqué par TH- » (Lapaire & Rotgé 1998 : 63). La réalisation fonctionnelle en pro-forme, qui est assez fréquente à l'écrit, peut s'expliquer par la nécessité de reprise, sans pour autant nommer le concept une nouvelle fois, celui-ci ayant déjà été construit dans le co-texte. On a donc une focalisation sur un concept qui ne requiert pas de répétition de la dénomination.

Tableau 6. Profils par ligne : pourcentage des fréquences relatives de *this* et de *that* pour chaque sous-corpus

Discourse	this proform	this determiner	this adverbial	that proform	that determiner	that adverbial	that complementizer	that relative pron
Oral medical	9	19	0	23	12	7	17	13
Oral science sci & tech (not med)	13	22	0	19	9	7	14	15
Whole oral corpus	9	15	0	25	9	7	22	13
Written medical	19	23	0	13	4	0	24	17
Written sci & tech (not med)	20	30	0	8	6	0	26	9
Written general	16	24	0	9	5	0	35	11

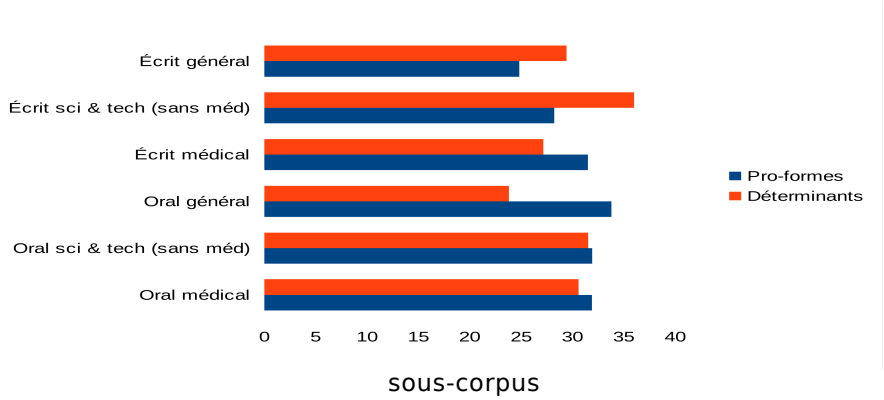
Les profils par ligne (tableau 6) permettent de mesurer la part que prend une forme par rapport aux autres au sein d'un même sous-corpus. On peut, pour un sous-corpus, comparer l'importance donnée à chaque forme et à sa catégorie grammaticale. On observe de nettes variations au sein de chaque corpus. La mise en regard des sous-corpus permet de comparer ces variations d'utilisation. Par exemple,

dans le sous-corpus oral médical, 9 % des *this* sont des *this* pro-forme, alors que ce ratio est de 19 % à l'écrit dans le domaine médical. On voit que *this* déterminant y est largement plus utilisé. Lors de l'analyse globale, les fréquences relatives de *this* étaient similaires entre oral et écrit. Mis en regard avec celles des autres formes pour un sous-corpus donné, on note que le poids relatif de *this* diffère. On observe en effet que, si on compte autant de *this* à l'écrit qu'à l'oral en anglais médical, le sous-corpus oral de ce domaine montre une sous-utilisation de *this* pro-forme par rapport à l'usage de toutes les autres formes de *this* et de *that*. Cette tendance est identique pour chaque sous-corpus oral. Inversement, *this* pro-forme joue un rôle majeur à l'écrit, ce qui rééquilibre la statistique globale. En ce qui concerne la forme *that,* sa part d'utilisation varie aussi selon les sous-corpus. *That* pro-forme joue un rôle significatif à l'oral, quel que soit le sous-corpus. *That* déterminant est plus utilisé à l'oral. Le complétif est plus utilisé à l'écrit qu'à l'oral, et son degré d'utilisation est plus grand en anglais général qu'en médical. Enfin, *that* relatif est plus fréquent dans le domaine spécialisé médical écrit que dans le domaine de l'anglais général à l'oral. Cette observation rejoint celle de Biber *et alii.* (1999 : 611) qui montrent que le relatif est moins présent en conversation (1 500 par million de mots) que dans les écrits scientifiques et universitaires (2 300 par million de mots). Ceci peut s'expliquer par le fait que l'écrit académique, proche du sous-corpus médical écrit, fait plus fréquemment appel à l'hypotaxe. Ce niveau de complexité syntaxique permet l'ajout d'informations pour le nom auquel renvoie *that.* Le concept en question reçoit des précisions concernant ses caractéristiques ou propriétés, ce qui est courant dans le discours écrit scientifique et universitaire.

Une dernière observation détaillée – toujours à partir des profils par lignes – concerne l'usage de la pro-forme par rapport à celui de la forme déterminante. L'objet d'une telle observation est de mesurer l'écart global entre les deux réalisations fonctionnelles. Il s'agit de voir si ces deux types de fonctionnement ont une régularité, ou bien si des spécificités liées à des corpus apparaissent. La somme des réalisations pour *this* et pour *that* permet de construire deux courbes et de constater la part relative de chaque catégorie grammaticale en fonction des sous-corpus (figure 4). Deux observations peuvent être faites. Premièrement, si on prend en compte le sous-corpus médical, celui-ci favorise légèrement les pro-formes à l'écrit, l'usage étant équivalent à l'oral. À l'inverse, les autres types de sous-corpus démontrent une variation de traitement entre les deux catégories, selon le mode oral ou écrit des sous-corpus. Les deux sous-corpus favorisent le déterminant à l'écrit, mais pas à l'oral. Deuxièmement, l'observation du mode oral révèle une utilisation équivalente des déterminants et pro-formes dans le domaine médical ainsi que dans le domaine scientifique et technologique, mais une proportion plus

forte de la pro-forme pour l'anglais général. Cet équilibre entre réalisation fonctionnelle déterminante et pro-forme peut provenir du type de discours. S'agissant de leçons ou de démonstrations, les situations expérimentales impliquent la mention des noms pour faire référence à des entités que le locuteur montre. Il s'agit de l'exophore telle qu'elle a été définie par Halliday et Hasan (1976 : 33). Dans ce type de situations, l'acte de monstration joue pleinement, en même temps que l'acte de dénomination. À l'inverse, la conversation en anglais général implique moins de dénomination. Une fois construit discursivement, un concept n'a plus besoin d'être nommé. Une reprise anaphorique par la pro-forme suffit pour maintenir la conversation. À titre de comparaison, Biber *et alii.* (1999 : 350) ont montré que l'usage de *that* en conversation est fait en pro-forme dans 70 % des cas, contre 15 % en déterminant. Si on observe le domaine de l'anglais général à l'oral, on note aussi que l'usage pro-forme domine, mais pas dans les mêmes proportions, sachant que cette mesure inclut l'usage de *this.*

Illustration 4. Pourcentage de pro-formes et de déterminants en fonction des



3.3. Évaluation statistique de la distribution

Abordons la question de l'indépendance des catégories de données du tableau servant à l'analyse. Il s'agit de savoir si la totalité des observations classées dans des catégories ont une variation aléatoire ou bien si elles sont causées par les catégories X et Y du tableau. En d'autres termes, la variation des fréquences d'une catégorie grammaticale de *that* ou de *this* est-elle liée à la spécificité du sous-corpus ? Le test du χ^2 permet de répondre à cette question. Le test a été effectué sur les données détaillées brutes (1) et normalisées (2) des formes grammaticales de *this* et de *that* en fonction des six sous-corpus (cf. tableaux 3 et 7). Il s'agit d'un calcul effectué sur l'ensemble des données. Son résultat mesure la différence entre les valeurs attendues et observées et compare cette différence avec ce qui serait attendu en cas de distribution due au hasard. Les résultats suivants sont obtenus :

1. X-squared = 2 610.465, df = 30, p-value < 2.2e-16. Without Yates' correction
2. X-squared = 16 532.18, df = 30, p-value < 2.2e-16. Without Yates' correction

Le résultat du χ^2 montre une corrélation entre les catégories sous-corpus et les catégories grammaticales de *this* et *that*. La probabilité de l'hypothèse nulle (le fait que les deux catégories soient totalement indépendantes l'une de l'autre et que les fréquences relevées ne soient que le fruit du hasard), mesurée par l'indicateur *p-value,* est très largement inférieure au seuil conventionnel de 0,05 (Gard 2008 *inter alia*). Le résultat du test semble indiquer un lien entre les deux catégories observées. Cependant, il ne renseigne pas sur l'ampleur de la différence entre les données du tableau. Pour avoir une idée de cette ampleur, il est nécessaire d'utiliser un test qui permet de voir dans quelle mesure l'information contenue dans une colonne contribue au résultat.

Tableau 7. Données brutes des formes dans le corpus ICE-GB pour le calcul du χ^2 (*this* adverbial est éliminé car n < 5)

Discourse	corpus subset size	this proform	this determiner	that proform	that determiner	that adverbial	that complementizer	that relative pron
Oral medical	15178	47	103	126	63	39	91	73
Oral science sci & tech (not m	20680	102	171	142	70	51	110	118
Oral general	637682	1750	2875	4949	1846	1451	4316	2617
Written medical	8431	35	43	23	7	0	45	31
Written sci & tech (not med)	38202	137	202	52	39	0	176	63
Written general	423581	1198	1816	677	406	8	2630	811

La mesure de Cramér V permet d'avoir des informations sur la contribution moyenne des lignes et des colonnes du tableau (Gard 2008). Pour Gard, le test de Cramér permet de réduire le rôle joué par l'impact de la taille de l'échantillon[5]. Le calcul s'effectue sur les données normalisées grâce à l'équation suivante sous R[6] :

$$V=\sqrt{\chi^2 \div \left(n*\left(k-1\right)\right)}$$

où n correspond à la somme totale des cellules du tableau et k correspond au plus petit nombre de colonnes ou de lignes. Sachant qu'un résultat de 100 % représenterait une association parfaite, le résultat obtenu ici montre une association relativement faible (14 %). La combinaison des mesures du χ^2 et du V de Cramér permet d'appréhender la corrélation des données croisées. Si le test d'indépendance du χ^2 montre une corrélation entre les variables des

[5] « Cramér V 'factors out' the size of the sample, and gives the 'average' contribution of rows and columns (that is, the catégories in the table and their respective observations in the rows and columns) to the final result ».

[6] Formule sous R : cv<-sqrt(chisq.test(thisthatnrmzd,correct=FALSE)$statistic/ (sum(thisthat)*min (dim(thisthat) – 1))) et résultat : X-squared = 0.1433692<. Explication donnée sur <http://www.biostat.wustl.edu/archives/html/s-news/2003-09/msg00185.html>. Consulté le 24 mai 2013.

sous-corpus et les variables relatives aux formes, le V de Cramér permet de prendre la mesure de la contribution moyenne des variables au résultat. Il s'agit ici d'un calcul prenant en compte l'ensemble des facteurs dans leur globalité et il est intéressant de mesurer les contributions individuelles aux résultats du χ^2. Cornillon *et alii* (2010 : 109) décrivent la procédure à suivre sous R avec la formule suivante :

```
round(100 * cthisthat$residuals^2 / cthisthat$stat, 1)
```

qui permet d'isoler les composantes individuelles du calcul du χ^2. « On peut étudier plus en détail cette liaison (avec le χ^2) en calculant les contributions à la statistique χ^2 observée. Les racines carrées des contributions sont dans l'objet residuals (du logiciel R). En divisant chaque terme par le total (*i.e.,* la valeur de χ^2 observée contenue dans la composante stat), on obtient un pourcentage » (Cornillon *et alii*. 2010).

Le tableau 8 fait la synthèse des contributions individuelles. Globalement, on trouve deux groupes de résultats. Le premier groupe concerne ceux qui contribuent faiblement au χ^2, par exemple, *this* déterminant à l'oral médical. Le second groupe concerne les combinaisons qui contribuent modérément à la non-indépendance des deux groupes de variables. Par exemple, *that* complétif est un levier important dans l'anglais écrit général.

Tableau 8. Calcul des contributions au résultat

	this.proform	this.determiner	that.proform	that.determiner	that.adverbial	that.complementizer	that.relative.pron
Oral medical	3.4	0.5	3.6	3	3.3	2	0
Oral sci & tech (not med)	0	0.1	0.1	0.2	2.2	5	0.6
Oral general	2.8	4	5.5	0.3	3.2	0	0
Written medical	3.3	0.3	2.1	3.1	6	0.6	1
Written sci & tech (not med)	4.2	4	6	0.7	4.8	1.3	1.4
Written general	0.6	0.4	4.7	1.1	4.7	9.3	0.6

Toujours selon Cornillon *et alii* (2010), « un retour aux données, via les résidus et,notamment, leurs signes », permet de voir quelles combinaisons favorisent l'hypothèse d'indépendance et avec quelle importance. Une valeur 0 indique une combinaison correspondant à une distribution attendue. Par exemple, les données du tableau 9 permettent de dire que *that* complétif est plus important qu'attendu à l'écrit. Inversement, le nombre d'occurrences de *that* en fonction adverbiale est beaucoup plus faible qu'attendu (statistiquement), en raison de sa très faible utilisation à l'écrit. On retrouve la très forte contribution du *that* pro-forme à l'anglais général et médical à l'oral. Sa contribution au sous-corpus science et technologie est nulle par rapport à ce qui serait attendu si l'hypothèse d'indépendance était vérifiée. De la même manière, *this* pro-forme est plus significatif à l'écrit dans le domaine médical et dans le domaine scientifique et technologique. *This* déterminant prend une plus grande place dans les productions écrites scientifique et technologique, et une place

beaucoup plus négligeable dans les productions relevant de l'oral général.

Tableau 9. Mesure de l'écart entre les contributions réelles et celles fondées sur l'hypothèse d'indépendance

	This proform	This determiner	That proform	That determiner	That adverbial	That complementizer	That relative pron.
Oral medical	-23.829	-9.349	24.443	22.35	23.37	-18.403	-0.619
Oral sci & tech (not med)	0.448	4.573	3.638	6.242	19.073	-28.688	9.671
Oral general	-21.432	-25.845	30.087	6.702	22.857	1.947	-1.794
Written medical	23.348	6.676	-18.524	-22.798	-31.534	10.16	13.06
Written sci & tech (not med)	26.206	25.561	-31.43	-11.072	-28.249	14.413	-14.982
Written general	9.565	8.044	-27.907	-13.273	-27.824	39.264	-10.283

Cette approche a permis un retour aux données. Les données du tableau montrent les résidus, c'est-à-dire les différences entre les fréquences relatives et les fréquences attendues. La section 4 poursuit cette étude des données en proposant l'analyse de textes provenant des deux sous-corpus médicaux oral et écrit.

4. Analyse de la distribution textuelle

Les analyses précédentes étaient centrées sur une approche statistique dont l'objectif était d'observer des tendances tout en assurant la validité de la distribution des données. Cette section consiste à effectuer un retour qualitatif sur certaines des données qui composent les sous-corpus médicaux oral et écrit.

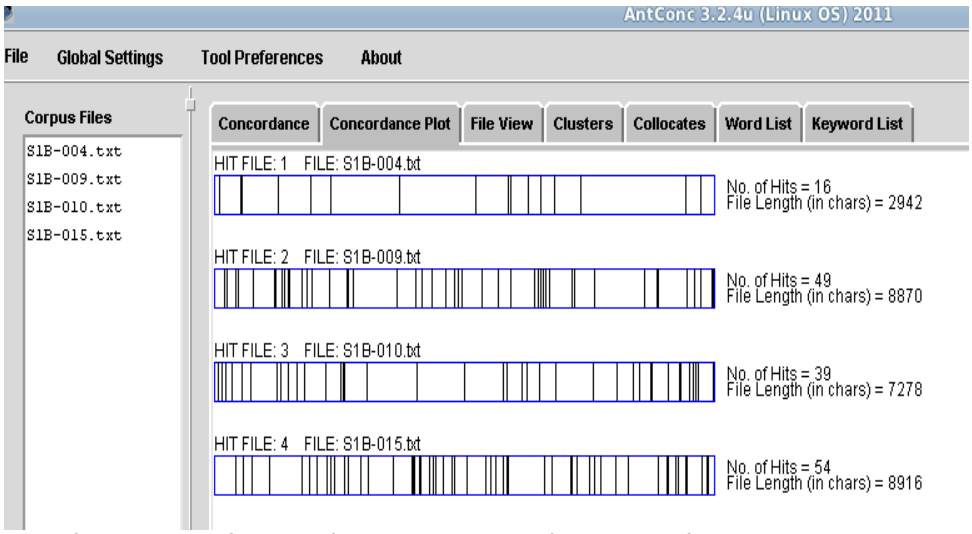
À partir de l'interface ICECUP ont été extraits quatre textes oraux (5 200 mots) et quatre textes écrits (9 000 mots). Ces textes ont été importés dans le concordancier AntConc[7] qui permet d'effectuer plusieurs types de calculs, dont les fréquences des formes et les fréquences de *clusters*, c'est-à-dire les fréquences de *n-grams* (séquence contiguë de n mots) incluant une forme pivot. En outre, on peut visualiser la distribution textuelle d'une forme, et savoir si on la trouve de manière régulière tout au long d'un texte ou plus regroupée en début, milieu ou fin de texte.

Si on s'intéresse à l'oral médical, et plus particulièrement aux *clusters 2-grams* incluant *that,* les plus fréquents sont *that's* (48 occurrences) et *than that* (10 occurrences). Il s'agit de l'usage de *that* en pro-forme, ce qui corrobore les observations faites plus haut. Pour *this,* le nombre d'occurrences dans les quatre textes est trop faible pour extraire des fréquences significatives, même si la pro-forme ressort statistiquement avec *this is* (2 occurrences). Il aurait été intéressant de procéder de la sorte avec des textes de l'anglais général afin de faire ressortir les *cluster 2-grams* les plus fréquents et ainsi de comparer avec le domaine médical. Cependant, cette partie qualitative se limite à la distribution textuelle dans les textes médicaux. La sélection de seulement quelques textes de l'anglais général s'appliquerait mécaniquement sur des champs particuliers, ce qui contredirait le concept même d'anglais général.

[7] <http://www.antlab.sci.waseda.ac.jp/antconc_index.html> consulté le 24 mai 2013

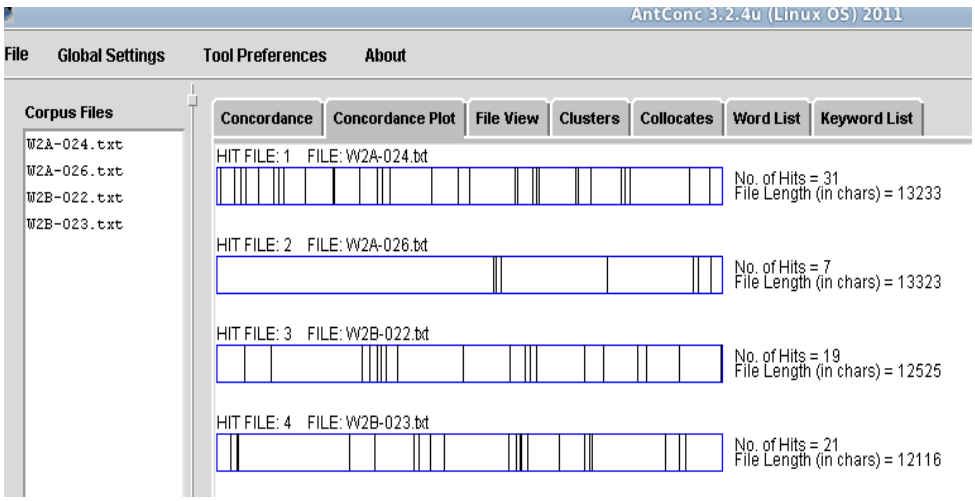
La mise en regard des graphes de distribution permet d'appréhender la distribution textuelle de la forme *that.* La figure 5 montre que, selon le locuteur, la densité textuelle à l'oral est variable. On note toutefois l'existence d'intervalles sans la forme après des périodes de densité forte. L'explication réside peut-être dans l'alternance de phases transitoires entre thématiques, avec des phases explicatives. Lors de phases explicatives ou descriptives dans le domaine médical, le locuteur est amené à devoir reprendre l'objet réel ou abstrait présenté dans son discours. Il tend donc à utiliser la pro-forme pour la mise en œuvre référentielle. D'autre part, le locuteur doit aussi expliciter des concepts et *that* complétif est adapté dans ce cas puisqu'il permet la subordination par rapport à des propositions englobantes telles que *you mean that*. L'étude statistique a montré que la pro-forme et le complétif étaient prépondérants dans le domaine médical à l'oral.

Illustration 5. Distribution textuelle de *that* dans des textes oraux médicaux



L'étude de l'écrit médical suit la même procédure. Les *clusters* les plus fréquents pour *that* sont *that the* (11), *that is/are* (7) et *that you* (5). Ce résultat correspond aux résultats statistiques montrant un usage important des *that* complétifs et relatifs au sein du sous-corpus médical. Pour *this,* les plus fréquents sont *this is* (7) et *this stage* (4). L'étude statistique montre une légère supériorité de *this* en pro-forme au sein du sous-corpus médical écrit et, malgré les faibles fréquences, on retrouve cette tendance dans l'échantillon. L'étude de la distribution textuelle de *that* dans des textes écrits médicaux (cf. figure 6) montre des écarts prononcés dans les usages. À l'écrit, le « which » concurrence la forme relative. Dans ce mode, la nécessité de reprise par le biais du *that* pro-forme semblerait moins forte. L'écrit médical appelle un usage endophorique de *that*. Or, les articles médicaux impliquent des développements cherchant à décrire et à mettre des concepts au premier plan, ce qui s'oppose à la valeur de rejet de *that* (Fraser & Joly, 1980 : 38). On retrouve aussi les phases de silence par rapport aux phases d'utilisation dense. Il y a moins d'opportunités d'utilisation du *that* au domaine médical écrit mais, lorsque celui-ci apparaît, les mécanismes de référence et de complexité syntaxique, avec le complétif ou relatif, sont pleinement exploités.

Illustration 6. Distribution textuelle de *that* dans des textes écrits médicaux



5. Conclusion

L'étude comparative globale révèle une différence d'usage entre *this* et *that*, que le domaine soit spécialisé ou non, et quelle que soit la spécialité. *This* est utilisé de manière équivalente à l'oral et à l'écrit, *that* prédomine à l'oral. Cependant, la présente étude, fondée sur la fonction grammaticale, montre que *this* et *that* ont un comportement spécifique dans les domaines spécialisés scientifiques. En effet, l'analyse détaillée des formes selon leur catégorie grammaticale montre une hétérogénéité au sein de l'apparente stabilité de *this.* Celui-ci se retrouve plus dans les textes scientifiques qu'en anglais général*. This* pro-forme est plus fréquent dans le mode mode écrit qu'à l'oral dans le domaine médical, par exemple. L'analyse détaillée de *that* révèle son rôle important en pro-forme à l'écrit, quel que soit le domaine, mais pas à l'oral. Le complétif suit une tendance inverse.

Le test d'indépendance des données effectué avec le calcul du χ^2 montre une dépendance de faible niveau entre les sous-corpus et les constituants grammaticaux. Ce résultat global faible cache des variations importantes lors de l'analyse des contributions individuelles à ce résultat. Celles-ci montrent que certaines formes comme le complétif verbal ont un rôle plus important que statistiquement attendu. La dernière partie de l'étude propose un retour sur les données en explorant la distribution textuelle d'un échantillon de textes oraux et écrits du domaine médical. Les rôles de certains constituants grammaticaux entrent en résonance avec l'étude statistique.

Cette étude contribue à l'effort de recherche visant à mettre à jour du rôle des différents constituants grammaticaux en fonction des discours. La démarche se positionne entre l'étude de la morphosyntaxe et des divers sens de marqueurs. Elle se fonde sur l'observation de l'évolution d'une forme en fonction de variations liées au contexte, donc au sens, et sur une analyse fonctionnelle des formes et de leurs caractéristiques morphosyntaxiques. Ce travail contribue de la sorte à l'inventaire des caractéristiques linguistiques propres à certains domaines spécialisés comme les sciences médicales, et il peut être articulé à leur enseignement et à leur apprentissage.

Remerciements

Nous remercions les participants au colloque GERAS pour leurs suggestions et remarques. Nos remerciements s'adressent aussi à Nicolas Ballier de l'Université Paris-Diderot, à Pascale Sébillot de l'IRISA/INSA de Rennes pour leur relecture et recommandations et à Geneviève Bordet de l'Université Paris-Diderot pour ses commentaires pertinents.

Références bibliographiques

AARTS, Bas, Gerald NELSON et Sean WALLIS. 2007. « Using fuzzy tree fragments to explore English grammar ». *English Today* 23/2, 27-31.

BIBER, Douglas, Stig JOHANSON, Geoffrey LEECH, Susan CONRAD et Edward FINEGAN. 1999. *Longman Grammar of Spoken and Written English*. Harlow : Longman.

BORDET, Geneviève. 2011. « *This* comme marqueur privilégié du genre : le cas des résumés de thèses ». *Discours* 9 [En ligne]. Consulté le 25 mai 2013. <http://discours.revues.org/8506>.

CARTER-THOMAS, Shirley et Elizabeth ROWLEY-JOLIVET. 2001. « Syntactic differences in oral and written scientific discourse: the role of information structure ». *ASp* [En ligne] 31-33. Consultation le 25 mai 2013. <http://asp.revues.org/1752>.

CORNILLON, Pierre André, Arnaud GUYADER, François HUSSON, Nicolas JÉGOU, Julie JOSSE, Maela KLOAREG, Eric MATZNER-LOBER et Laurent ROUVIÈRE. 2010. *Statistiques avec R*. 2e édition augmentée. Rennes : Presses Universitaires de Rennes.

CORNISH, Francis. 1999. *Anaphora, Discourse, and Understanding. Evidence from English and French*. Oxford : Oxford University Press.

COTTE, Pierre. 1993. « De l'étymologie à l'énonciation ». *Travaux de linguistique et de philologie* 31, 43–89.

FRASER, Thomas et André JOLY. 1979. « Le système de la deixis - Esquisse d'une théorie d'expression en anglais ». *Modèles Linguistiques* 1/2, 97–157.

FRASER, Thomas et André JOLY. 1980. « Le système de la deixis (2) : Endophore et cohésion discursive en anglais ». *Modèles Linguistiques* 2/2, 22–51.

GARD, B. Jenset. 2008. « Basic statistics for corpus linguistics ». Université de Bergen. Consulté le 25 mai 2013. <http://folk.uib.no/gje037/statTutorialR.pdf>.

GAILLAT, Thomas. 2013. « *This* and *that* in native and learner English: from typology of use to tagset characterisation ». In S. GRANGER, G. GILQUIN et F. MEUNIER (dir.) (2013) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use - Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain.

GRIES, Stefan Th. 2010. « Useful statistics for corpus linguistics ». In AQUILINO S. et A. MOISÉS (dir.), *A Mosaic of Corpus Linguistics: Selected Approaches*. Peter Lang : Francfort-sur-le-Main, 269–291.

HALLIDAY, Michael A. K. et Ruqaiya HASAN. 1976. *Cohesion in English*. English Language Series. Harlow : Pearson Education Limited.

HUDDLESTON, Rodney et Geoffrey K. PULLUM. 2002. *The Cambridge Grammar of the English Language*. Beccles, Suffolk : Cambridge University Press.

HYLAND, Ken. 2010. « Constructing proximity: Relating to readers in popular and professional science ». *Journal of English for Academic Purposes* 9/2, 116–127.

KLEIBER, Georges. 1992. « Anaphore-deixis : deux approches concurrentes ». In MOREL, M.-A. et L. DANON-BOILEAU, *La Deixis*. Paris : Presses Universitaires de France, 613–626.

LAPAIRE, Jean-Rémi et Wilfrid ROTGÉ. 1998. *Linguistique et grammaire de l'anglais*. Toulouse : Presses Universitaires du Mirail.

NELSON, Gerald, Sean WALLIS et Bas AARTS. 1998. « The British component of the International Corpus of English (ICE-GB) and ICECUP Software (CD-ROM) ». Londres. Consulté le 25 mai 2013. <http://www.ucl.ac.uk/english-usage/projects/ice-gb/>.

NELSON, Gerald, Sean WALLIS et Bas AARTS. 2002. « Exploring natural language: Working with the British component of the International Corpus of English ». Amsterdam : John Benjamins Publishing Co.

PARKINSON, Jean et Ralph ADENDORFF. 2004. « The use of popular science articles in teaching scientific literacy ». *English for Specific Purposes* 23/4, 379–396.

PETIT, Michel. 2010. « Le discours spécialisé et le spécialisé du discours : repères pour l'analyse du discours en anglais de spécialité ». *E-rea* [En ligne] 8/1. Consulté le 25 mai 2013. <http://erea.revues.org/1400>.

QUIRK, Randolph, Geoffrey LEECH et Jan SVARTVIK. 1985. *A Grammar of Contemporary English*. Londres, Beccles et Colchester : Longman.