La notion de collocation fondamentale : une étude de corpus

Veronica Benigno, Olivier Kraif, Francis Grossmann, Antonino Velez

LA NOTION DE COLLOCATION FONDAMENTALE
UNE ETUDE DE CORPUS

Veronica Benigno[1], Olivier Kraif[1], Francis Grossmann[1], Antonino Velez[2]

[1] Univ. Grenoble Alpes, [2] Università degli studi di Palermo
veronicabenigno@hotmail.com, {olivier.kraif, francis.grossmann}@u-grenoble3.fr,
antonino.velez@unipa.it

Résumé

Dans cette étude les 'collocations fondamentales' sont envisagées comme des unités polylexicales significatives (unies par des liens collocationnels) fréquentes (dans l'usage) ou non fréquentes (lorsqu'elles sont pertinentes pour la communication) qui représentent pour les locuteurs natifs les contextes les plus essentiels d'un mot donné. Nous avons extrait du corpus *frWaC* (Baroni et al., 2010) environ 20 000 associations à partir de dix mots pivot issus su *Dictionnaire Fondamental* de Gougenheim (1971); puis, au moyen de la fréquence et de mesures d'association, nous avons sélectionné un échantillon d'environ 400 associations candidates au statut de collocations fondamentales et demandé à 90 locuteurs natifs du français de sélectionner celles qui leur apparaissaient essentielles pour la communication. L'objectif était : a) de développer une méthode pour le repérage automatique des collocations fondamentales ; et b) d'évaluer, à l'aide de l'intuition de locuteurs natifs, la validité de l'échantillon automatiquement constitué et de comprendre de quoi dépendait l'assignation du caractère fondamental de la part des enquêtés. Nous avons constaté que lorsque la fréquence ne jouait pas de rôle, pour les enquêtés, dans l'attribution du caractère fondamental la cause était à chercher dans le figement. D'après les résultats obtenus, nous avons pu confirmer l'intuition gougenheimienne sur l'équation *fondamental = fréquence et/ou utilité communicative*.

Mots clés : vocabulaire fondamental, collocation fondamentale, corpus, fréquence, utilité communicative figement

Abstract

The present study investigates 'core collocations', i.e. frequent or available (i.e. essential for accomplishing basic communicative tasks) word combinations consisting of two lexemes which yield a significant (collocational) relation and which represent the most basic co-occurrences of a word. We extracted from the web-crawled corpus *frWaC* (Baroni et al., 2010) a sample of around 20 000 collocations, using frequency and associative measures. Then, we selected the top 400 collocations by frequency and/or associative measures and asked 90 native speakers to indicate which ones they considered to be more essential. The purpose of the study was twofold: a) developing a valid method to automatically extract core collocations; and b) determining, by means of a comparison between native intuition and statistical data, the validity of the sample we automatically compiled and the

reason why some combinations were regarded as more central by the subjects. The study showed that what is core is not only a matter of frequency, but it also depends on units' fixedness. Our findings support Gougenheim's intuition that coreness depends on frequency and/or communicative usefulness.

Keywords : core vocabulary, core collocation, corpus, frequency, communicative usefulness, fixedness

1. Introduction

Dans le passé, de nombreuses études ont élaboré des listes de fréquence appelées listes de base, listes réduites, simplifiées, élémentaires, etc. dans lesquelles les mots du vocabulaire fondamental sont énumérés et décrits. Aujourd'hui, ces études apparaissent incomplètes pour avoir négligé l'analyse des associations typiques et fréquentes qui opèrent en tant qu'unités lexico-grammaticales. Les mots de base sont les plus connus chez les locuteurs natifs car ils leur permettent d'accomplir les actes du quotidien, mais ils n'existent pas en tant qu'unités isolées : bien au contraire, leur sens se réalise au sein de collocations, d'où la nécessité d'étudier les associations qu'ils privilégient et les contextes d'occurrences où ils apparaissent le plus fréquemment.

Du fait de leur caractère relativement arbitraire et idiosyncratique, les collocations représentent un obstacle majeur dans l'apprentissage et dans l'enseignement des langues étrangères. Cela explique pourquoi nombre d'études sont publiées régulièrement sur le sujet, et ce par des linguistes aux spécialisations les plus variées. Les collocations ont fait l'objet de nombreuses recherches qui ont tenté de saisir leurs propriétés (parmi des synthèses récentes, citons Haussmann : 1979 ; Heid : 1994 ; Cowie : 1998 ; Grossmann, Tutin : 2003 ; Bartsch : 2004 ; Carter, Schmitt : 2004; Tutin : 2010), de comprendre la façon dont elles doivent être traitées dans les dictionnaires traditionnels ou électroniques (Lo Cascio, 1999), de proposer des théories nouvelles qui expliquent le fonctionnement du lexique mental et de l'apprentissage lexical (Wray : 2000). Cependant, étant donné la multiplicité des angles d'approche du phénomène, il est bien difficile d'aboutir à une définition faisant consensus. Quant aux ressources existantes pour l'apprentissage et la didactique de la langue étrangère, bien qu'elles aient déjà permis un important pas en avant dans le traitement des collocations, elles sont encore loin d'être suffisantes.

Si l'inventaire du français fondamental s'était ouvert au syntagme (cf. Galisson, 1971), l'unité centrale prise en compte restait le mot. L'analyse conduite dans la présente étude élargit la notion de « vocabulaire fondamental » (Gougenheim et al. : 1964) à la dimension syntagmatique et propose une méthode pour identifier et définir ce que nous proposons d'appeler « collocation fondamentale ». Dans la présente étude, nous avons comparé les collocations fondamentales repérées automatiquement et des collocations fondamentales sélectionnées par les locuteurs natifs. Nous avons ainsi essayé de répondre à trois questions principales : quelle est la corrélation entre fréquence et jugement sur le caractère fondamental des

associations tel qu'il est évalué par les locuteurs natifs ? De quel facteur autre que la fréquence dépend l'attribution du caractère fondamental de la part des locuteurs natifs ? Quelle est l'utilité de la fréquence, des mesures d'association et de la dispersion pour le repérage des collocations fondamentales ?

2. Repères théoriques

2.1. Le vocabulaire fondamental

Les mentions de nombreuses recherches sur le lexique ont montré l'importance de disposer d'un vocabulaire fondamental[1], c'est-à-dire d'un ensemble de mots réputés essentiels pour penser ou exprimer les concepts de base dans une langue. Qu'appelle-t-on « vocabulaire fondamental » d'une langue ? Il s'agit du noyau lexical d'une langue, le vocabulaire dont chaque locuteur dispose pour ses actes communicatifs élémentaires et quotidiens (Carter, 1998).

Qu'il s'agisse de listes rédigées sur la base de considérations d'auteurs sur le caractère essentiel du vocabulaire (par exemple le *Basic English* d'Ogden en 1930) ou de listes rédigées à l'aide de mesures statistiques comme les comptages de fréquence (c'est le cas de la liste *French Word Book* de Vander Beke en 1929), les premières listes de vocabulaire simplifiées ou réduites sont nées en raison des besoins de l'enseignement de la langue étrangère ; ce n'est que par la suite qu'elles se sont élargies à d'autres domaines, qu'elles ont visé d'autres objectifs, tels que la mesure de la lisibilité des textes ou l'étude du style d'un auteur.

En français, l'ouvrage de référence du vocabulaire fondamental est né d'une initiative de l'U.N.E.S.C.O. visant à diffuser les grandes langues de civilisation. *Le Français Elémentaire* (1954) représente la liste de fréquence la plus connue et la plus significative, élaborée d'après de grandes enquêtes sur la langue orale (plus récemment, on y a intégré des données concernant la langue écrite). L'ouvrage inclut les mots les plus fréquents, ainsi que des « mots disponibles » : il s'agit de mots moins fréquents mais pertinents par rapport à leur domaine sémantique spécifique et qui sont considérés comme essentiels par les locuteurs natifs. Des mots comme *fourchette* et *dent* mériteraient d'être inclus dans une liste fondamentale car, comme l'explique Gougenheim *et al.* (1964, p. 145), ils sont « […] cependant des mots usuels et utiles ». Ainsi, déjà dans les années 50, cette étude de pionnier révélait que le caractère fondamental n'était pas à attribuer uniquement à la fréquence. Cette démarche novatrice nous semble pouvoir être reprise et étendue aux collocations : en combinant, tout comme Gougenheim, les analyses statistiques et l'intuition des sujets natifs, il est possible, selon nous, de mieux comprendre ce dont dépend l'attribution du caractère fondamental, la fréquence d'emploi n'étant pas toujours un facteur stable. L'organisation, en 2005, d'un colloque à l'ENS de Lyon titré

[1] Pour des précisions sur les débats concernant le français fondamental et un approfondissement des aspects théoriques, voir Benigno et al., 2015.

« Français Fondamental, corpus oraux, contenus d'enseignement, 50 ans de travaux et d'enjeux » montre la pertinence actuelle du travail de l'équipe de Gougenheim et son importance dans le domaine didactique (Cortier & Parpette, 2006).

2.3. Définition de travail des collocations

La définition de travail de la collocation que nous adoptons s'inspire des conceptions de Heid (1994), Mel'čuk (1998), Grossmann et Tutin (2002), ou encore Hausmann (2007) : il s'agit d'« une séquence polylexicale qui actualise un mot dans une unité sémantico-syntaxique typique, et qui se caractérise par le sémantisme transparent de la base et le sémantisme restreint du collocatif ». Cette définition met en évidence que :

a. La collocation est conçue comme une séquence résultant de l'association entre deux ou plusieurs éléments *lexicaux* :

> If we accept that collocations are combinations of two lexemes, and that they together form a linguistic object which needs to be described in a dictionary, it is consequent to also accept that in total there are three kinds of objects to be described in a collocational dictionary: the base (in our case the noun), the collocate (the Qi$A) verb) and the collocation as a whole; if there are no space restrictions, it makes sense to ensure that collocations have lemma status in a dictionary (Heid, 1998; 304; cf. aussi Heid 1994:241).

Autrement dit, la collocation n'est pas une association dont l'un des composants est un élément grammatical : ainsi, une expression comme *par hasard,* qui associe un élément grammatical et un élément lexical, ne sera pas considérée comme une collocation au sens strict, bien qu'elle soit une association importante d'un point de vue didactique.

b. La collocation s'actualise dans une unité sémantico-syntaxique typique, représentative de la façon dont un concept est mis en contexte par les locuteurs d'une langue. Comme l'explique bien Sinclair (1991, p. 44), à un sens correspond une structure, ainsi le sens influence la structure et vice-versa En anglais, par exemple, le verbe *to decline* dans le sens de « refuser » se présente toujours au passé simple, comme dans l'exemple *He declined his invitation* (Sinclair, 1991).

c. La collocation se caractérise par le sémantisme transparent de la base, car la base garde toujours son sens plein, tandis que le sens du collocatif peut être plus ou moins opaque. Considérons les collocations *commettre un crime* et *caresser l'espoir :* dans les deux cas, le substantif en fonction de base est transparent. C'est pourquoi la collocation est plus facilement comprise au décodage pour un étranger.

d. La collocation se caractérise par le sémantisme restreint du collocatif. Le collocatif peut avoir à la fois un sens littéral ou non littéral, mais il a toujours un sémantisme restreint : son association avec la base est exclusive ou quasi-exclusive

dans un sens spécifique (le « meaning by collocation » firthien). Autrement dit, la commutabilité synonymique du collocatif est exclusive ou limitée : en anglais, il est possible de dire *to lose control*, mais il n'est pas possible de dire *to miss control*. Cela dérive du fait que la collocation est soumise à des contraintes de nature sémantique imposées par l'arbitraire de l'usage linguistique :[2]

> (…) nous pouvons décrire la collocation (exemples: passer un examen, un célibataire endurci, grièvement blessé, une bouffée de colère) comme la combinaison phraséologique (codée en langue) d'une base (examen, célibataire, blessé, colère) et d'un collocatif (passer, endurci, grièvement, bouffée). La base est un mot (plus précisément l'acception d'un mot, appelée aussi «lexie») que le locuteur choisit librement parce qu'il est définissable, traduisible et apprenable sans le collocatif. Le collocatif est un mot (ou l'acception d'un mot) que le locuteur sélectionne en fonction de la base parce qu'il n'est pas définissable, traduisible ou apprenable sans la base (Hausmann & Blumenthal, 2006 :4)

Haussmann (1984) fait une distinction entre éléments lexicaux « autosémantiques », éléments autonomes au niveau sémantique car leur sens est indépendant de leur contexte syntagmatique ; et élément lexicaux « synsémantiques », éléments dépendants sémantiquement de leur contexte syntagmatique, c'est-à-dire de leur base. Dans les associations libres, les deux constituants sont autosémantiques. Dans les locutions, les deux composants sont synsémantiques parce qu'ils se sélectionnent réciproquement. Dans les collocations, seulement le collocatif est synsémantique : en association, la base garde son sens, tandis que le collocatif se définit par rapport à la base.

3. Constitution de l'échantillon, corpus et outil d'analyse

Le choix de repérer les collocations à partir de leurs bases est opéré dans la plupart des ouvrages lexicographiques, étant donné que, comme l'explique Hausmann (1979), l'utilisateur d'un dictionnaire cherche son collocatif à partir d'une base connue. Le groupe de mots pivots qui représentent la « tête » ou « base » des collocations dans notre étude est constitué de dix substantifs : *colloque, conférence, congrès, conversation, débat, fête, interview, rencontre, réunion, séminaire.* Le substantif représente généralement (voir Grossmann, Tutin, 2003 pour des exceptions) la tête lexicale dans le couple base-collocatif pour de nombreuses raisons. En premier lieu, le substantif a un caractère référentiel et dénotatif plus

---

[2] Selon Nesselhauf (2005), Hausmann a le mérite d'avoir été le premier à reconnaitre que les composants de la collocation ne possèdent pas tous le même statut, ce qui avait était négligé par les statisticiens : « En effet, dans la collocation *célibataire endurci*, le signifié de la base (*célibataire*) est autonome. La base n'a pas besoin du collocatif (*endurci*) pour être clairement définie. Il en va tout autrement pour le collocatif qui ne réalise pleinement son signifié qu'en association avec une base (*célibataire*, *pécheur*, *âme,* etc.) » (Hausmann 1979, p. 191, cité par Nesselhauf).

précis que les adjectifs et les verbes (au moins dans la langue générale). En outre, comme le montrent les études en psycholinguistique sur l'accès au lexique mental, l'accès aux noms s'effectue plus rapidement que pour les verbes (voir par exemple Ellis, 1997). Les dix substantifs sont extraits des articles « Sociabilité », « Compagnie » et « Relation » dans le *Thésaurus Larousse* de Péchoin (1991) et ils sont fondamentaux, c'est-à-dire fréquents ou disponibles selon le *Dictionnaire Fondamental de la langue française* de Gougenheim (1971).

Les associations produites à partir des dix mots pivots choisis ont été extraites du corpus *frWaC* (Baroni et al. : 2010), un corpus de plus d'un milliard de mots constitué de textes issus du Web, récupérés par des procédures automatiques. La constitution de *frWaC* s'est déroulée en deux étapes fondamentales, la sélection des mots clés (« seeds » en anglais) et le *crawling,* d'une part, et d'autre part, le nettoyage et l'annotation. Les deux principaux critères qui ont guidé notre choix sont la très grande taille du corpus et sa couverture de la langue générale : en effet, les auteurs montrent qu'en termes de couverture, la version anglaise du *frWaC*, constituée selon la même procédure, contient beaucoup plus de types que le BNC (environ 5 fois plus pour les noms et les adjectifs et 3 fois plus pour les verbes – en ne retenant que les types apparaissant au moins 20 fois). La notion de « langue générale » n'est pas aisément saisissable car elle suppose que le corpus qui la représente soit très vaste et le plus varié possible en termes de genres textuels. Bien que le *frWaC* ne constitue pas un échantillon contrôlé en terme de genres, de types de discours, de domaines sémantiques, etc. il est suffisamment vaste et hétérogène pour présenter une grande variété de types de textes ressortissant de la langue générale : il est constitué de forums, de blogs, d'article de presse, d'articles d'encyclopédie, de romans, d'écrits académiques, etc. ; en outre, il présente une population de locuteurs très variée en termes d'âge et de niveau d'éducation.

Le corpus *frWaC*, lemmatisé et annoté morphosyntaxiquement avec Treetagger (Schmid, 1994), a été exploré à l'aide de scripts Perl développés par nous (Kraif, 2011). La taille de l'empan compte 4 cooccurrents à gauche et 4 à droite pour l'examen des concordances, 50 caractères sont affichés à gauche et à droite du pivot dans la sortie KWIC. Nous précisons que la ponctuation est comptée dans l'empan. Parfois, des contextes plus larges sont nécessaires pour que le collocatif soit repéré. Cependant, notre corpus est très grand et la perte de quelques résultats significatifs est secondaire : ainsi, nous préférons réduire le bruit découlant d'un empan trop large. L'exécution des scripts produit en sortie deux fichiers au format .TXT qui représentent l'output de l´interrogation : un concordancier et une liste de fréquence accompagnée de mesures statistiques.

- Le concordancier affiche les lignes de concordances pour chaque cooccurrence du mot pivot. Les sources web d'où elles sont extraites, c'est-à-dire l'URL, sont aussi indiquées. Nous utilisons les concordances pour dériver les patterns sémantiques et syntaxiques caractérisant les cooccurrences extraites.

- La liste de fréquence (absolue) des associations ainsi que les statistiques suivantes : le nombre total de tous les couples de cooccurrents dans le corpus, y compris ceux de la base et du collocatif ; les mesures d'association telles que le *log-likelihood ratio* (Dunning, 1993), le *t-score*, le *z-score* et l'information mutuelle spécifique notée IM (Evert, 2008). Ces différentes mesures d'association permettent de normaliser la fréquence absolue des cooccurrences en tenant compte des fréquences des occurrences prises indépendamment. En effet, la simple fréquence de cooccurrence risque d'une part d'ignorer des associations peu fréquentes mais significatives, comme par exemple *réunion consultative*, et d'autre part, de faire ressortir des combinaisons libres ultra fréquentes mais non significatives comme *la réunion* (déterminant + nom). Une mesure répandue telle que l'IM permet de repérer les associations peu fréquentes (McEnery et al. : 2006, p. 57) mais dont les composants ont une association très forte en termes d'information partagée, comme *nez aquilin* et *au fur et à mesure*. Nous évaluerons plus loin les profils comparés de ces différentes mesures.

Enfin, une dernière colonne indique la « dispersion », pour mesurer la répartition d'une cooccurrence dans les documents du corpus. Comme il s'agit d'une collection de textes issus du Web, nous avons calculé la dispersion comme le nombre de noms de domaines différents (p. ex. www.lemonde.fr, www.libe.fr, www.le-figaro.fr) dans lesquels une cooccurrence se présente. Si un phénomène est toujours observé sur un même site, sa dispersion vaut 1 : il peut donc s'agir d'un phénomène local et sans portée générale, malgré une haute fréquence. A l'opposé, un nombre élevé de domaines différents est une bonne indication de la fiabilité de l'observation statistique, et est le signe d'un bon candidat pour les collocations fondamentales.

4. Méthodologie

Pour chaque mot pivot analysé, on n'a retenu que les 2 000 associations les plus fréquentes. Ce nombre a été établi après un contrôle manuel des associations retenues par des différents seuils statistiques. On a constaté que l'extraction de 2 000 cooccurrents était largement suffisante pour qu'aucune cooccurrence significative n'échappe au calcul et pour limiter le bruit. Ce nombre se réduit après un premier filtrage automatique (correction des erreurs de lemmatisation, élimination des catégories vides, des signes de ponctuation, des cooccurrents peu dispersés, etc.) à environ 1 000 cooccurrents, et il diminue ultérieurement (jusqu'à environ 300-400 cooccurrents) après un dernier filtrage manuel des sorties (élimination d'entités nommées, de collocations incomplètes, etc.). En outre, on a pris en compte l'orthographe exacte du mot, en tenant compte de la casse. Si le pivot est *porte*, alors *Porte* est considérée comme une autre forme et n'est pas comptabilisée (sauf si elle est en majuscule au début de la phrase). La normalisation des majuscules (par ex. en début de phrase) et des majuscules emphatiques (*Ministère* et *ministère*) n'étant pas correctement effectuée par *TreeTagger,* il a fallu ajouter un deuxième mécanisme de

lemmatisation « forcée » pour regrouper ces formes. Tous ces paramètres ont permis de réduire le bruit et d'affiner la recherche selon des critères établis *a priori*.

Ainsi, environ 20 000 unités polylexicales ont été repérées dans le corpus *frWaC* (Baroni et al. : 2010), au moyen de scripts Perl, à partir de dix substantifs pivots représentant la « tête » ou « base » de l'unité. Ensuite, des opérations à la fois automatiques et manuelles nous ont permis de réduire de façon importante la liste de candidats (élimination des entités nommées, des unités peu dispersées, etc.). Nous ne retenons, pour certains mots pivots, que 300 à 400 cooccurrents sur les 2 000 initiaux. Ensuite, afin de constituer notre échantillon de collocations fondamentales, nous réduisons l'échantillon à 40-50 associations par pivot par un filtrage basé sur nos différentes mesures d'association.

4.1. *L'échantillon de collocations fondamentales*

Le recours aux mesures d'association présente l'intérêt d'écarter les mauvais candidats dérivant du simple calcul de fréquence, tels que certaines colligations ou combinaisons d'unités très fréquentes. Le fait d'utiliser une combinaison de mesures permet de profiter de leur complémentarité, certaines étant plus sensibles aux associations de basse fréquence (comme l'IM), et d'autres à celles de haute fréquence (comme le *t-score* ou le *log-likelihood*) (Evert, 2008). Afin de constituer notre échantillon de collocations fondamentales, nous avons donc filtré nos sorties en fonction de l'ensemble des mesures d'association : nous avons retenu les cooccurrents qui étaient significatifs selon au moins une des mesures associatives choisies (l'information mutuelle, le log-likelihood, le t-score et le z-score). Il était nécessaire d'établir un critère pour le filtrage : pour chaque mesure (sauf IM), nous avons retenu les 50 meilleurs candidats. L'établissement de ce seuil arbitraire a été guidé par une analyse qualitative des résultats extraits. On constate, en effet, l'absence de cooccurrents significatifs au-delà de ce seuil. Nous avons noté que ce seuil était parfois trop tolérant, mais nous avons préféré augmenter le bruit et ne pas avoir à exclure de cooccurrents potentiellement significatifs. Pour l'IM, nous avons utilisé un autre seuil de significativité, en ne retenant que les associations avec un score au moins égal à 3 (Church & Hanks, 1990).

Lors de cette phase de filtrage, la consultation des concordances des résultats extraits nous a permis d'éliminer les collocations invalides encore présentes à ce stade.

4.2. *Le test soumis auprès de locuteurs natifs*

Ensuite, un test a été soumis auprès de locuteurs natifs pour évaluer l'échantillon constitué et comprendre quel type d'unités était perçu comme étant fondamental. Nous avons interrogé 90 locuteurs de nationalité française, suisse ou belge (dont le niveau d'instruction, l'âge et le domaine professionnel étaient variés – voir Annexe A). La plupart des locuteurs ont répondu à une annonce qui leur expliquait l'objectif et l'utilité du test et qui a été postée dans le réseau social Web « couchsurfing » (www.couchsurfing.com). Le test s'est déroulé de la façon

suivante : avant de commencer le test, les locuteurs ont été invités à lire les instructions, qui leur fournissaient une explication générale du concept d'« association fondamentale », avec des exemples extraits du corpus (voir Annexe B). Après la lecture des instructions, les locuteurs étaient censés commencer le test (voir Annexe C). Les associations fréquentes et celles moins fréquentes extraites à l'aide de l'outil d'exploration du corpus leur étaient présentées dans une feuille de calcul, et on leur demandait d'exprimer un jugement sur leur caractère fondamental. Ils devaient marquer d'une croix les associations qui leur semblaient les plus familières et les plus essentielles pour la communication.

Nous voudrions souligner que le test envisagé pour la présente recherche a été conçu pour éviter de surcharger les locuteurs avec des explications et des tests trop techniques: le jugement de « fondamental » devait être rendu sur une base intuitive.

Un total d'environ 450 associations[3] a été réparti entre 90 répondants divisés en 3 groupes de 30 locuteurs chacun, avec pour le groupe 1 les pivots *conférence, débat, rencontre*, pour le groupe 2 les pivots *colloque, fête, réunion* et enfin, pour le groupe 3, les pivots *congrès, séminaire, interview, conversation*. Bien que certains pivots puissent apparaitre comme plus fondamentaux que d'autres, ils font tous partie du vocabulaire de base selon le *Dictionnaire Fondamental* de Gougenheim (1971). Les associations ont été présentées dans leur structure lexico-syntaxique d'occurrence la plus fréquente : par exemple, étant donné que *conférence* se présente toujours au pluriel en association avec *cycle*, la forme suivante a été choisie : *un cycle de conférences*. La même chose vaut pour l'association entre *conférence* et *s'intituler* : ce dernier se présentant le plus souvent dans notre corpus sous la forme verbale du présent, nous avons présenté aux répondants l'association *la conférence s'intitule*. Le même principe vaut pour l'alternance article défini/indéfini dans l'association de *animation* et de *conférence* : la forme la plus fréquente dans l'usage, *l'animation de la conférence*, a été choisie. Quant à l'ordre de présentation des associations, il a été choisi au hasard.

5. Résultats

Etant donné que les mesures d'association nous informent sur la force d'association entre les composants de l'unité, mais non sur l'utilité communicative de l'unité polylexicale, l'échantillon des associations extraites a ensuite été soumis au jugement des locuteurs natifs. Le dépouillement des réponses des enquêtés qui ont complété leurs tests entièrement et lisiblement nous a permis de tirer des résultats significatifs, présentés ci-après.

---

[3] Notons que le nombre d'unités polylexicales soumises au jugement des locuteurs était variable selon les résultats issus de l'extraction automatique.

5.1 Une corrélation positive non systématique

Le premier résultat de notre étude est la présence d'une corrélation faiblement positive, mais comme nous allons le voir, non systématique, entre fréquence des associations extraites du corpus et score obtenu par les locuteurs. Cela veut dire que les associations les plus fréquentes ne sont pas *toujours* les plus sélectionnées. Si une corrélation est présente, elle est affaiblie par la présence de points singuliers qui s'écartent de la norme. Dans le tableau 1, nous représentons les scores du coefficient de corrélation de Pearson obtenus par les dix mots pivots :

PIVOTS	COEFF		PIVOTS	COEFF
colloque	-0,19		Fête	0,16
conférence	0,43		Interview	0,05
congrès	0,08		Rencontre	- 0,11
conversation	0,27		Réunion	0,29
débat	0,30		séminaire	0,29

Tableau 1 : Coefficient de corrélation (Pearson) entre la fréquence et le nombre de votes positifs, pour les dix pivots

Le coefficient de corrélation de Pearson permet de mesurer numériquement l'écart de la distribution des points par rapport à la droite de régression. Si le caractère fondamental des associations dépendait linéairement de la fréquence, on trouverait une valeur proche de 1. Mais, dans nos observations nous avons toujours obtenu des valeurs inférieures à 0,5, avec un maximum de 0,43 pour *conférence*. Dans notre exemple, si on avait une corrélation parfaite entre les deux variables, les associations jugées comme les plus fondamentales par les locuteurs natifs devraient *toujours* être les plus fréquentes, ce qui n'est pas le cas pour certaines associations.

Pour *conférence*, par exemple, nous avons remarqué l'existence d'une corrélation faible, mais positive : au fur et à mesure que la fréquence de l'association augmente, le score attribué par les locuteurs à l'association croît également. Cependant, quelques associations s'éloignent du « centre » de la distribution, car elles représentent une « déviation » dans l'évolution moyenne de la corrélation entre fréquence et caractère fondamental assigné par les interviewés. Comme nous le montrons dans le Tableau 2, les données sont très dispersées, car certaines associations sont plus à l'écart par rapport aux autres. Les points singuliers les plus extrêmes, de haute fréquence et de score faible (en gras), sont *la conférence de rentrée* (1053, 7%), *le président de la conférence* (1203, 17%) et *avoir une conférence* (8069, 20%) ; tandis que des points singuliers de basse fréquence et de score élevé (en italique) sont *une conférence intergouvernementale* (492, 40%), *une conférence téléphonique* (458, 83%) et *le compte-rendu de la conférence* (436, 70%). Il s'agit d'unités qui sont très fréquentes mais peu retenues par les locuteurs (par exemple, *avoir une conférence*), ou à l'inverse, d'unités peu fréquentes mais souvent retenues (par exemple, *une conférence téléphonique*).

ASSOCIATIONS	FREQUENCE	SCORE NATIFS
une * de presse	14207	97%
un maître de *s	14106	87%
être à une *	10418	57%
*avoir une **	*8069*	*20%*
organiser une *	5589	70%
une * internationale	3729	73%
donner une *	3487	83%
un cycle de *s	3111	40%
la salle de *	3045	93%
le thème de la *	2478	77%
au cours de la *	2439	63%
tenir une *	2045	57%
le programme de la *	1749	50%
*faire une **	*1674*	*37%*
animer une*	1331	73%
participer à une *	1233	73%
*le président de la **	*1203*	*17%*
à l'occasion de la *	1166	53%
assister à une *	1130	93%
*la * de rentrée*	*1053*	*7%*
une * annuelle	1043	43%
l'animation de la *	809	27%
une * de consensus	790	7%
une * épiscopale	609	17%
prononcer une*	608	13%
la * s'intitule ...	595	23%
une * mensuelle	569	27%
une * ministérielle	524	40%
*une * intergouvernementale*	*492*	*40%*
*une * téléphonique*	*459*	*83%*
*le compte-rendu de la **	*436*	*70%*
la tenue de la *	382	10%
une * inaugurale	248	23%
une * tripartite	147	17%
une * introductive	102	17%

Tableau 2 : Associations avec conférence *triées par fréquence décroissante et quelques points singuliers (en italique).*

Le fait que la fréquence ne puisse pas être considérée comme un facteur absolu, dont dépendrait le jugement des locuteurs de ce qui serait ou non fondamental, apparaît encore plus évident si nous analysons le mot *colloque* (Tableau 3), car la corrélation entre les deux variables est faiblement négative. Le mot *colloque*, de la même façon que le mot *conférence*, présente des associations très fréquentes mais peu sélectionnées, et à l'inverse, des associations moins fréquentes mais souvent retenues. Dans le Tableau 3, nous remarquons des points de haute fréquence et de score faible plus nets : *un colloque scientifique* (1003, 3%), *le thème du colloque* (1215, 17%) et *les communications du colloque* (1605, 17%) ; et des points de basse fréquence et de score élevé: *un colloque pluridisciplinaire* (116,

73%), *la clôture du colloque* (234, 87%) et *le compte-rendu du colloque* (255, 90%) :

ASSOCIATIONS	FREQUENCE	SCORE NATIFS
les actes du *	7056	47%
organiser un *	5772	43%
un * international	5056	27%
être à un *	4565	23%
avoir un *	3220	73%
*les communications du **	*1605*	*17%*
l'organisation du *	1451	37%
*le thème du **	*1215*	*17%*
tenir un *	1188	80%
un * national	1007	83%
*un * scientifique*	*1003*	*3%*
le programme du *	886	23%
une journée de *	886	80%
à l'occasion du *	790	33%
participer au *	699	73%
dans le cadre du *	686	63%
le * a lieu …	673	63%
un * consacré à ...	664	77%
un * d'étude	601	63%
présenter un *	587	43%
faire un *	562	50%
un * annuel	559	60%
la participation au *	535	43%
l'ouverture du *	530	40%
un * intitulé X	515	20%
les participants au *	513	37%
l'intervention au *	495	80%
réunir un *	491	47%
la présentation du *	399	50%
le prochain *	374	73%
la contribution au *	288	10%
le * se déroule...	282	7%
*le compte-rendu du **	*255*	*90%*
*la clôture du **	*234*	*87%*
l'organisateur du *	221	73%
*un * interdisciplinaire*	*183*	*73%*
un * singulier	133	47%
*un * pluridisciplinaire*	*116*	*73%*
un * co-organisé	78	50%

Tableau 3 : Pivot colloque *: liste des associations triées par fréquence décroissante et quelques points singuliers (en italique).*

Pour conclure, nous pouvons affirmer que, globalement, il existe une corrélation linéaire positive entre la fréquence et le caractère fondamental, mais que la présence de points singuliers qui s'écartent de la norme affaiblit cette corrélation et montre qu'elle n'est pas systématique. Les locuteurs natifs ont parfois ignoré certaines associations très fréquentes, tandis qu'ils ont d'autres fois retenu des

associations plus rares. Cette observation confirme l'hypothèse selon laquelle ce qui est fondamental ne ressort pas uniquement du facteur de la fréquence. De quel facteur dépend alors l'attribution du caractère fondamental ?

Le second résultat de notre étude est obtenu à travers une analyse plus attentive de nos données afin de comprendre la raison pour laquelle certaines associations s'écartent de la norme. Nous avons remarqué que le degré de figement jouait un rôle dans l'attribution du caractère fondamental, au moins pour les points singuliers les plus nets. Ainsi, une association peu fréquente est quand même retenue lorsqu'elle comporte un certain degré de figement ; à l'inverse, une association très fréquente est exclue lorsqu'elle est totalement libre.

6. Discussion

Considérons à présent la valeur des mesures d'association pour le repérage des collocations fondamentales. Pour ce faire, nous devons prendre en compte la spécificité de notre corpus issu du Web. Bien qu'il soit relativement adéquat dans le cadre de notre analyse en ce qui concerne sa taille et sa représentativité de la langue générale, il faut reconnaitre qu'il comporte des biais, notamment à cause de la présence de pages Web qui se répètent et qui proposent des contenus similaires. En outre, l'empan temporel assez restreint du processus de crawling (10 jours) confère une saillance artificielle à des faits d'actualité, qui occupent une part non négligeable des contenus engrangés. Parmi les collocatifs extraits, on trouve des intrus anormalement fréquents : ce sont en général des entités nommées, des noms propres, des toponymes, qui n'entrent pas dans notre définition des collocatifs (ces entités nommées, comme par exemple les *fêtes panathénées*, ne sont pas toujours marquées d'une majuscule). Pour atténuer ces biais, nous avons croisé la fréquence avec les mesures d'association et la valeur de dispersion. Ici, il est utile d'examiner précisément comment ces différentes mesures ont pu contribuer au repérage des collocations fondamentales.

Les mesures d'**IM** et de ***t-score*** nous semblent les moins utiles au repérage des collocations fondamentales. En effet, nous avons remarqué qu'elles repèrent davantage des associations très rares ou celles dont la dispersion possède une valeur basse, du bruit et des entités nommées. Par exemple, pour le pivot *rencontre* (avant le nettoyage des sorties), l'IM extrait un seul collocatif significatif, l'adjectif *nationale*, tandis que le *t-score* n'en extrait aucun. Les deux mesures extraient, par contre, des collocatifs tels que *jacques*, *gilles et goody* qui ne présentent aucun intérêt en tant que collocatifs de *rencontre*.

Le ***log-likelihood*** favorise les paires fréquentes et n'extrait pas les associations peu dispersées, et il récupère moins de bruit et d'entités nommées que les deux mesures précédentes. Cependant, cette mesure n'est pas non plus capable d'éliminer complètement le bruit, car elle ramène un certain nombre d'associations anormalement fréquentes liés aux actualités correspondant à la période de

constitution du corpus (par exemple, pour le pivot *rencontre* : *paris, françois, udf-modem, bayrou, clae*).

La **fréquence** et le ***z-score*** apparaissent – étonnamment ! – comme étant des mesures plus précises pour le repérage des collocations fondamentales. Pour *rencontre*, par exemple, les deux mesures ne récupèrent qu'une seule collocation invalide (avec *premier* qui est le plus souvent suivie de *ministre)*.

La **dispersion**, parfois négligée dans les études sur les collocations, se révèle être une mesure cruciale pour le repérage des collocations fondamentales. Etant donné que nous nous sommes intéressés à un phénomène central dans la langue, il était essentiel d'éliminer les valeurs de fréquence liées à des singularités du corpus. Certaines associations comme *fêtes panathénées* ou *fête (du) Beaujolais* obtiennent de très bons scores, mais sont peu dispersées. La suppression de ces associations a été permise grâce à l'élimination de toutes les cooccurrences ayant une valeur de dispersion égale ou inférieure à 50 sources. Ce seuil a été établi d'après le constat que, avec un seuil de dispersion inférieur à 50, presque aucune cooccurrence n'est correcte ou considérée comme acceptable. Parmi les opérations de réduction du bruit présent dans le corpus, celle-ci est la plus consistante car elle permet de réduire la liste des sorties d'environ 1 000 résultats, ce qui représente un chiffre conséquent.

Nous n'avons pas effectué une mesure précise de corrélation pour ces différentes mesures, dans la mesure où nos observations manuelles nous ont montré que la fréquence constituait le meilleur critère pour presque toutes les unités considérées.

Ces observations confirment, comme l'avait noté Evert (2008), qu'il n'y a pas une mesure d'association qui serait meilleure qu'une autre dans l'absolu : tout dépend du profil des unités visées (phénomènes de basse, moyenne ou haute fréquence) des filtres appliqués (p.ex. élimination des colligations, dispersion) et de la structure du corpus. Dans la mesure où l'on cherche des phénomènes qui se situent plutôt dans les moyennes et hautes fréquences, nous avons constaté avec une certaine surprise qu'un simple filtrage des colligations sur une base grammaticale (élimination des mots fonctionnels) permet d'éliminer une grande partie du bruit. Le bruit qui reste est constitutif du corpus, et ne peut être éliminé grâce à des mesures plus sophistiquées telles que le *loglike* : il s'agit de phénomènes anormalement fréquents lié à l'empan temporel trop étroit. Si l'aspiration du corpus s'était effectuée sur des périodes différentes et distantes, afin d'assurer une certaine dispersion temporelle des phénomènes, ce bruit aurait été naturellement filtré : au plan du repérage automatique, des progrès restent à faire, mais surtout du côté de l'échantillonnage et de l'équilibrage d'un grand corpus de référence.

7. Conclusion

Dans cette étude, nous avons cherché à caractériser empiriquement un type d'expression encore assez peu étudié : les collocations fondamentales. Pour ce faire

nous avons abordé le phénomène sous deux angles complémentaires : celui des mesures objectives basées sur la fréquence et les indices d'association, et celui du jugement subjectif de pertinence. Nous avons dans un premier temps réalisé une extraction automatique des collocations candidates à partir d'un vaste corpus, le *frWaC*. Ces collocations ont ensuite été soumises à 90 locuteurs chargés de les annoter, afin d'en signaler le caractère essentiel et utile pour la communication usuelle. Nos observations montrent une certaine corrélation entre l'attribution subjective de ce caractère et la fréquence absolue, sauf dans le cas d'expressions situées aux deux extrêmes du figement : des expressions perçues comme fréquentes mais complètement libres étant écartées, alors que des expressions plus rares mais plus figées sont retenues. Ces observations ont confirmé l'hypothèse initiale selon laquelle les collocations fondamentales sont des associations fréquentes ou non fréquentes qui se caractérisent par un certain degré de figement.

Cette étude nous a permis, par ailleurs, de comparer la pertinence de différentes mesures d'association pour l'extraction automatique. Il nous est apparu que le *z-score* ou le *loglike* étaient assez adaptés à un tel corpus de grand volume – ce qui est conforme aux observations effectuées dans la littérature. De façon plus inattendue, nous avons constaté que même la fréquence absolue peut-être un critère utilisable, du moment qu'on élimine les associations avec les mots grammaticaux et que ces indices sont croisés avec la mesure de dispersion, qui garantit la généralité de l'expression candidate. Un échantillonnage du corpus garantissant une meilleure dispersion temporelle aurait par ailleurs permis d'obtenir de meilleurs résultats. D'après les résultats obtenus, nous avons pu confirmer l'intuition gougenheimienne sur l'équation *fondamental = fréquence et/ou utilité communicative*. Ainsi, les recherches en didactique du lexique et en lexicographie ne devraient pas tomber dans le piège de traduire le sens de « fondamental » *uniquement* par « fréquent », car cela signifierait fonder l'enseignement du vocabulaire et de la phraséologie seulement sur des facteurs quantitatifs, insuffisants en tant que tels.

Références

Bally, C. (1951). *Traité de stylistique française*. Paris : Klincksieck.

Baroni, M., Bernardini, S., Ferraresi, A., Picci, G., (2010). Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation. In Xiao R. (Ed.), *Using Corpora in Contrastive and Translation Studies*. Newcastle : Cambridge Scholars Publishing.

Bartsch, S. (2004). *Structural and functional properties of collocations in English*, Thèse de doctorat. Tübingen: Gunter Narr Verlag.

Benigno, V. (2007). Il vocabolario di base : tratti costitutivi, rilevanza cognitiva e acquisizione in italiano L2. In Lo Cascio V. (Ed.), *Parole in rete : apprendimento e teoria nell'era elettronica.* Novara : Utet-Università, pp. 151-174.

Benigno, V., Grossmann, F. & Kraif, O. (2015). Les collocations fondamentales : une piste pour l'apprentissage lexical, *Revue française de linguistique appliquée,* vol. XX, 81-96.

Brent, W. (2009). Meaning-last vocabulary acquisition and collocational productivity » dans Fitzpatrick T., Barfield A. (Ed.), *Lexical Processing in Second Language Learners : Papers and Perspectives in Honour of Paul Meara (Second Language Acquisition)*. Bristol : Multilingual Matters, p. 128-140.

Carter, R. (1998). *Vocabulary. Applied Linguistic Perspectives*. London & New York: Routledge, 2e éd.

Carter, R., Schmitt, N. (2004). Formulaic sequences in action. An introduction. In Schmitt N. (Ed.), *Formulaic sequences. Acquisition, processing and use*. Amsterdam: John Benjamins Publishing Company, pp. 1-22.

Cortier, C., Parpette, C. (Eds) (2006). De quelques enjeux et usages historiques du Français fondamental. *Documents pour l'histoire du français langue étrangère ou seconde*, 36

Cowie, A. P. (1998). *Phraseology. Theory, Analysis, and Applications*. Oxford: Clarendon Press.

Church, K. W., Hanks, P. (1990). Word association norms, mutual information,and lexicography. *Computational Linguistics*, vol. 16, p. 22-29.

Dunning, T. (1993). Accurate Methods for the Statistic of Surprise and Coincidence. *Computational Linguistics*, 19(1), pp. 61-76.

Ellis, R. (1997). *Second language acquisition*, Oxford: Oxford University Press

Evert, S. (2008). Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*. Berlin : Mouton de Gruyter, pp 1212-1248.

Galisson R. (1971). *Inventaire thématique et syntagmatique du français fondamental.* Paris : Hachette-Larousse, Coll. Le Français dans le Monde/ BELC.

Gougenheim, G., Michéa, R., Rivenc, P., Sauvageot, A. (1964). *L'élaboration du français fondamental (1er degré)*. Nouvelle édition refondue et augmentée. Paris : Didier.

Grossmann, F., Tutin, A. (2002). Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*, VII-1, pp. 7-25.

Grossmann, F., Tutin, A. (Ed.) (2003). *Les collocations : analyse et traitement*, Amsterdam : De Werelt.

Hausmann, F., J. (1979). Un dictionnaire des collocations est-il possible ? ». *Travaux de linguistique et de littérature XVII* (1), Strasbourg, p. 187-195.

Hausmann, F. J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen franzosischer Wortverbindungen, dans *Praxis des neusprachlichen Unterrichts* 31, p. 395-406

Hausmann, F., J., Blumenthal P. (2006). Présentation : collocations, corpus, dictionnaires, *Langue française* 2006/2 (n° 150), p. 3-13.

Hausmann, F., J. (2007). DIE KOLLOKATIONEN IM RAHMEN DER PHRASEOLOGIE-SYSTEMATISCHE UND HISTORISCHE DARSTELLUNG, in *Zeitschrift für Anglistik und Amerikanistik*, Band 55.3, Collocation and creativity,Verlag Königshausen & Neumann, p. 217-234

Heid, U. (1998). Towards a corpus-based dictionary of German noun-verb collocations, in *Proceedings of the EURALEX International Congress 1998*, Liège.

Heid, U. (1994). On Ways Words Work Together - Topics in Lexical Combinatorics, in W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, P. Vossen (Eds.), *Euralex 1994, Proceedings*, Amsterdam, 1994, p. 226-257.

Hoey, M. (2005). *Lexical priming. A new theory of words and language*. London: Routledge.

Kraif, O. (2011). Les concordances pour l'observation des corpus : utilité, outillage, utilisabilité. In Jean Chuquet (sous la dir. de), *Le langage et ses niveaux d'analyse.* Rennes : Presses universitaires de Rennes (PUR), chap. 4, pp. 67-80.

Lo Cascio, V. (1999). Standardisation and Collocation. In Telen M., Lewandowska-Tomaszczyk B. (Ed.), *Translation and meaning- part 5.* Maastricht: Hoogschool Zuyd, pp. 23-38.

Mel'čuk, I. (1998). Collocations and Lexical Functions » dans Cowie A. P. (Ed.), *Phraseology. Theory, Analysis, and Applications.* Oxford: Clarendon Press, p. 23-53.

Manning, C. D., Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

McEnery, T., Xiao, R., Tono, Y. (2006). *Corpus-based language studies. An advanced resource book*. London: Routledge.

Nesselhauf, N. (2005). *Collocations in a learner corpus - Studies in corpus linguistics*. Amsterdam et Philadelphia: John Benjamins Publishing Company

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Sinclair, J. (1991). *Corpus, concordances, collocation*. Oxford: Oxford University Press.

Sinclair, J., Jones, S., Daley, R. (2004). *English collocation studies : the OSTI report*. London & New York: Continuum.

Tutin, A. (2010) *Sens et combinatoire lexicale : de la langue au discours*, Synthèse d'Habilitation à diriger des recherches, Université Stendhal Grenoble 3.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Annexe A – Provenance, âge et profession des 90 informateurs

Locuteur	Sexe	Age	Profession ou domaine d'étude	National-ité
1	F	30	Doctorante en linguistique	Française
2	M	36	Traductuer et enseignant de français	Française
3	F	26	Assistante aux ventes	Française
4	M	27	Enseignant secondaire - secteur Arts	Belge
5	M	27	Ingégnieur informatique	Belge
6	M	37	Enseignant universitaire - Faculté des Lettres et Sciences humaines	Suisse
7	M	43	Agent de maitrise dans une usine	Française
8	M	21	Baccalauréat en enseignement primaire	Belge
9	M	22	Master en informatique	Française
10	M	35	Enseignant universitaire - Faculté des Lettres et Sciences humaines	Suisse
11	M	33	Technicien - licence DUT	Française
12	M	25	Doctorant en physique	Française
13	F	31	Ergothérapiste/Rééducatrice en service de soins à domicile	Française
14	F	25	Juriste en droit de l'environnement marine	Française
15	M	30	Directeur artistique	Française
16	F	32	Enseignante secondaire - secteur Arts	Belge
17	M	28	Expert informatique	Française
18	F	28	Assistante aux ventes	Française
19	F	24	Master en muséo-expographie	Française
20	M	27	Infographiste	Française
21	F	21	Ethologue équin	Française
22	F	24	Médiatrice culturelle	Française
23	F	28	Opératrice dans le sectuer du tourisme	Française
24	M	24	Pompier	Française
25	F	29	Doctorante en linguistique	Française
26	M	34	Mécanicien motocycles	Française
27	M	24	Doctorant en chimie	Française
28	M	28	Consultant informatique	Française
29	F	34	Conseillère à l'emploi	Française
30	M	23	Etudiant d'école d'ingénieur en génie industriel	Française
31	M	28	Etudiant en urbanisme	Française
32	M	63	Enseignant secondaire - secteur Lettres	Française
33	F	31	Etudiante en master en management des organisations culturelles	Française
34	F	53	Secrétaire de Mairie	Française
35	F	29	Journaliste	Française
36	F	33	Archiviste	Française
37	M	49	Enseignant universitaire - secteur information et communication	Française

38	F	26	Archiviste	Française
39	M	45	Agent immobilier	Française
40	F	36	Enseignante universitaire - Dept. Sciences du Language	Française
41	F	48	Enseignante secondaire - anglais	Française
42	F	33	Enseignante secondaire - anglais	Française
43	M	31	Doctorant en linguistique	Française
44	F	36	Enseignante secondaire	Française
45	F	29	Doctorante en linguistique	Française
46	F	24	Chef de projet transversal a l`Institut Francais de Roumanie	Française
47	F	26	Interprète en langue des signes française	Française
48	F	27	Etudiante de licence en chinois	Française
49	F	27	Infirmière	Française
50	M	25	Doctorant en ingénierie	Française
51	M	40	Cadre supérieur dans la distribution textile	Française
52	M	28	Responsable de gestion de projets	Française
53	M	24	Ingégnieur civil	Française
54	F	52	Expert d'accompagnement socio-professionnel des jeunes demandeurs d’emploi	Française
55	F	25	Ingénieure mécanique/énergétique	Française
56	M	29	Employé dans le secteur du développement durable	Française
57	F	41	Enseignante universitaire - Dept. Sciences du Language	Française
58	M	28	Enseignant secondaire	Française
59	F	27	Employée dans le secteur ressources humaines	Française
60	M	31	Ecrivain	Française
61	F	53	Traductrice, interprete et professeur de FLE	Française
62	M	27	Etudiant en informatique	Française
63	F	37	Doctorante en linguistique et enseignante universitaire	Belge
64	F	25	Docteur en pharmacie	Française
65	F	26	Employée (secteur SIG - Système d'Information Géographique-)	Française
66	F	26	Etudiante de master en politiques sociales	Française
67	M	22	Etudiant de licence en Géographie et Aménagement du Territoire	Française
68	F	34	Enseignante secondaire - français	Française
69	F	43	Responsable Fédérale - Union Démocratique Bretonne	Française
70	F	27	Enseignante dans l'école primaire	Française
71	M	37	Enseignant secondaire (latin, religion et culture antique)	Française
72	F	36	Logopède et enseignante secondaire - français	Française
73	M	33	Enseignant secondaire	Française
74	F	20	Employée - secteur tourisme	Française
75	F	46	Enseignante universitaire - Dept. Sciences du Language	Française
76	M	26	Architecte	Française
77	M	59	Employé - secteur commercial	Française

78	F	34	Doctorante en informatique	Française
79	M	40	Cadre supérieur dans la distribution textile	Française
80	F	30	Responsable de management qualité	Française
81	F	27	Employée - secteur communication	Française
82	M	31	Logiciel informatique et formateur	Française
83	F	25	Etudiante de master en sociologie	Française
84	M	25	Salarié en management des risques	Française
85	F	24	Etudiante de master en histoire	Française
86	F	31	Employée - secteur administration de la recherche universitaire	Française
87	F	52	Directrice Générale IRFA	Française
88	M	28	Chef d'entreprise	Française
89	F	46	Etudiante de master en litérature	Française
90	F	30	Enviromentaliste	Française

Annexe B – Instructions du test soumis auprès des locuteurs natifs

POURQUOI CE TEST ?

Dans la liste qui suit, sont présentés, pour un mot donné (qu'on appellera « base »), d'autres mots qui s'y associent en contexte : par exemple, la « base » *séminaire* peut s'associer avec les autres mots suivants (parmi de nombreux d'autres) :

1. **organiser** un *séminaire*
Ex. L'association organise un séminaire annuel sur le racisme.
2. *séminaire* **de formation**
Ex. L'AFNIC intervient régulièrement dans des séminaires de formation de Télécom Paris.
3. *séminaire* **résidentiel**
Ex. De 1986 à 1990, le CEFAG se compose de cinq séminaires résidentiels.

Chacun d'entre vous, lorsqu'il entend un mot comme *séminaire* est capable de lui associer spontanément d'autres mots, parce que ses emplois sont fréquents ou importants dans la vie courante : parmi ces associations privilégiées, on trouve par exemple *organiser* un *séminaire*. En revanche, il vous viendrait probablement moins facilement à l'esprit une association comme *séminaire résidentiel*, parce que l'on n'en a besoin que rarement.

Nous vous demandons de repérer les associations privilégiées et de laisser de côté les associations que vous jugerez moins fondamentales, du type *séminaire résidentiel*.

COMMENT EXECUTER LE TEST ?

Il vous est demandé de cocher avec un X, les mots associés permettant de construire des associations privilégiées ou fondamentales. Le test ne prend que quelques minutes et doit suivre trois règles de base :

1. Réfléchir quelques secondes à chaque proposition de mot associé

2. Imaginer la façon dont la « base » et l'autre mot s'associent. Par exemple, à partir de *séminaire* et *recherche*, il est possible de former l'association *séminaire de recherche.*

3. Mettre une croix dans la colonne de droite chaque fois que vous jugerez que l'association entre les deux mots est « fondamentale » (c'est-à-dire immédiatement identifiable, parce que fréquente ou importante pour les besoins de la vie courante).

Annexe C – Pivot "Conversation" - Le test soumis auprès des locuteurs natifs

BASE: conversation	**Note: La touche * indique la position de la base, avant ou après le mot associé**
Mots associés	**Fondamental?**
à la fin de la *	
alimenter la *	
animer la *	
au cours d'une *	
au détour d'une *	
au fil de la *	
avoir une *	
commencer la *	
continuer la *	
écouter la *	
en pleine *	
engager la *	
enregistrer une *	
entamer la *	
entendre la *	
interrompre une *	
la * courante	
la * orale	
la * quotidienne	
la suite de la *	
lancer une *	
l'enregistrement de la *	
mener une *	
mettre fin à la *	
participer à la *	
poursuivre la *	
reprendre la *	
suivre la *	
surprendre une *	
tenir une *	
tourner la *	
un sujet de *	
une * au téléphone	
une * avec le jury	
une * écrite	
une * en langue X	
une * intéressante	
une * privée	
une * téléphonique	
une courte *	
une longue *	
une minute de *	
une nouvelle *	
une simple *	
Le test est terminé. Est-ce qu'il y a d'autres associations fondamentales qui vous viennent à l'esprit? Si oui, indiquez-les dans la ligne suivante.	