CHACQFAM : une base de données renseignant l'âge d'acquisition estimé et la familiarité pour 1225 mots monosyllabiques et bisyllabiques du Français

Christian Lachaud

NOTE THEORIQUE

Laboratoire de Psycholinguistique Expérimentale,

Faculté de Psychologie et des Sciences de l'Éducation,

Université de Genève

CHACQFAM : une base de données renseignant l'âge d'acquisition estimé et la familiarité pour 1225 mots monosyllabiques et bisyllabiques du Français

Dr. Christian Michel LACHAUD

Université de Genève

FPSE, Laboratoire de Psycholinguistique Expérimentale

Bd. du Pont d'Arve, 40

CH 1205 Genève, GE, Suisse

RESUME

L'âge d'acquisition et la familiarité d'un mot sont des facteurs décisifs pour l'accès au lexique, en production comme en perception. Pour favoriser les recherches sur les mécanismes du traitement lexical en Français, une base de données lexicales a été constituée pour un corpus de 1225 mots monosyllabiques et bisyllabiques du Français. Cet article décrit la méthode utilisée pour le recueil des données, l'information brute obtenue, la procédure pour traiter cette information brute, et le contenu de la base de données CHACQFAM obtenu après traitement de l'information brute, ainsi que les procédures de validation de ce contenu. CHACQFAM est disponible gratuitement sur le site Internet « http://psycholinguistique.unige.ch/ ».

Mots-clé :

CHACQFAM, âge d'acquisition, familiarité, base de données lexicales, Français.

CHACQFAM: a lexical data base for the estimated age of acquisition and familiarity of 1225 monosyllabic and bisyllabic French words

SUMMARY

Age of acquisition of a word and familiarity are important factors for the lexical processing, in production as in perception. To help developing research on the mechanisms underlying the lexical processing in French, a lexical data base was built for a corpus of 1225 monosyllabic and disyllabic French words. This article describes the method used to collect the data, the rough information obtained with the survey, explains the method that was used to process the rough information, describes the content of the lexical data base CHACQFAM, obtained after the rough data has been processed, and describes the validation procedure of its content. CHACQFAM is made freely available to researchers in an electronic format, from the website "http://psycholinguistique.unige.ch/".

Key words:

CHACQFAM, age of acquisition, familiarity, lexical data base, French.

INTRODUCTION

Les bases de données lexicales représentent un outil de travail essentiel pour l'étude des mécanismes du traitement du langage. En psycholinguistique, deux bases de données lexicales du Français, BRULEX (Content, Mousty, et Radeau, 1990) et LEXIQUE (New, Pallier, Ferrand, et Matos, 2001), permettent de disposer d'indices descriptifs formels des mots, tels que le nombre de lettres et de phonèmes constituant un mot, la fréquence avec laquelle ce mot est utilisé dans la langue, le genre du mot ou sa catégorie grammaticale. Par contre, ces deux banques de données ne renferment pas d'indices psychologiques, tels que la familiarité d'un individu pour un mot, la valence d'imagerie du mot, ni d'indices développementaux tel que l'âge d'acquisition d'un mot. Ces dernières années ont donc vu apparaître des bases de données lexicales complémentaires, proposant de l'information pour de telles variables (Alario et Ferrand, 1999 ; Bonin, Méot, Aubert, Malardier, Niedenthal, et Capelle-Toczek, 2003 ; Bonin, Peereman, Malardier, Méot, et Chalard, 2003 ; Chalard, Bonin, Méot, Boyer, et Fayol, 2003 ; Desrochers et Bergeron, 2000 ; Ferrand, Grainger, et New, 2003). Ces banques de données psychologiques et développementales sont tout aussi essentielles que les deux premières pour l'étude des mécanismes du traitement du langage. Cependant, la quantité de mots actuellement couverte par ces bases complémentaires s'avère être relativement limitée. C'est dans l'idée de contribuer à l'accroissement des moyens de recherche à disposition de la communauté psycholinguistique, qu'une nouvelle base de données lexicales pour la langue française, intitulée CHACQFAM (1), est proposée, et présentée à travers cet article.

---

[1] Ce nom a été choisi relativement au nom du pays depuis lequel la base de données a été constituée (Confédération Helvétique), et à l'information qu'elle contient (L'âge d'ACQuisition et la FAMiliarité).

CHACQFAM renseigne deux variables pour 1225 mots du français, l'âge d'acquisition et la familiarité, dont les scores ont été estimés par des adultes. La base de données elle-même est mise à disposition du public dans un format électronique à partir du site Internet « http://psycholinguistique.unige.ch/ », d'où elle peut être téléchargée.

L'âge d'acquisition d'un mot (AoA) et la familiarité pour un mot (Fam) sont deux facteurs connus pour influencer significativement et de façon robuste le traitement lexical, en production comme en perception, et en modalité visuelle comme en modalité auditive. De plus, les effets de ces deux variables semblent être universels, puisqu'ils ont été retrouvés dans différentes langues. Ils se traduisent comportementalement par une augmentation du temps de traitement d'un mot avec l'âge auquel l'individu a appris ce mot (AoA), et par une augmentation du temps de traitement d'un mot avec la diminution de la familiarité de l'individu avec ce mot (Fam).

Le lecteur est invité à consulter l'abondante littérature qui existe sur la question pour les détails techniques des expériences ayant montré un effet d'AoA et de Fam, ainsi que pour les hypothèses proposées afin de rendre compte des mécanismes psychologiques sous-jacents (par exemple, pour l'AoA : Alario, Ferrand, Laganaro, New, Frauenfelder, et Segui, 2004 ; Barry et Gerhand, 2003 ; Barry, Hirsh, Johnston, et Williams, 2001 ; Barry, Morrison, et Ellis, 1997 ; Baumeister, 1985 ; Bird, Franklin, et Howard, 2001 ; Bonin, Chalard, Meot, et Fayol, 2002 ; Bonin, Fayol, et Chalard, 2001 ; Brown et Watson, 1987 ; Brysbaert, 1996 ; Brysbaert, Lange, et Van Wijnendaele, 2000 ; Chalard et al., 2003 ; Ellis et Morrison, 1998 ; Garlock et Walley, 2001 ; Gerhand et Barry, 1998 ; Gerhand et Barry, 1999 a ; Gerhand et Barry, 1999 b ; Izura et Ellis, 2002 ; Juhasz, 2005 ; Kremin, Hamerel, Dordain, De Wilde, et Perrier, 2000 ; Lyons, Teer, et Rubenstein, 1978 ; Meschyan et Hernandez, 2002 ; Monaghan et Ellis, 2002 ; Morrison et Ellis, 1995 ; Morrison et Ellis, 2000 ; Morrison, Hirsh, Chappell, et Ellis, 2002 ; Morrison, Hirsh, et Duggan, 2003. Pour la familiarité : Alario et al., 2004 ; Connine,

Mullennix, Shernoff, et Yelen, 1990 ; Hirsh et Funnell, 1995 ; Kremin, Hamerel, Dordain, De Wilde, et Perrier, 2000).

En Français, CHACQFAM complète un patrimoine de 4 banques de données de même nature (Alario et al., 1999 ; Bonin et al., 2003 ; Chalard et al., 2003 ; Ferrand et al., 2003).

Quantitativement, par son volume lexical relativement important (1225 mots, contre 403 pour la base de Ferrand et collègues, 400 pour la base d'Alario et Ferrand, 299 pour la base de Bonin et collègues, et 230 pour celle de Chalard et collègues), et par son faible chevauchement avec les bases précédemment citées (86 mots en commun avec la base de Ferrand et collègues (soit 7% des mots de CHACQFAM), 49 mots avec la base d'Alario et Ferrand (4%), 59 mots avec la base de Bonin et collègues (5%), 36 mots avec la base de Chalard et collègues (3%)), CHACQFAM représente un outil de travail non négligeable.

Qualitativement, CHACQFAM fournit une information précise et fiable, qui n'est pas entièrement similaire à celle proposée dans les 4 bases déjà existantes. Concernant l'âge d'acquisition tout d'abord, la précision offerte par la méthode de recueil spécifiquement développée est optimale, puisque les réponses sont données avec un pas annuel plutôt que par tranches d'âge (de 2 ans pour la base de Ferrand et collègues, de 3 ans pour celles d'Alario et Ferrand et de Bonin et collègues). Ceci permet incidemment d'éviter l'emploi d'une limite supérieure dans l'échelle utilisée (âges de 13 ans et plus groupés dans une même catégorie dans la base de Ferrand et collègues, de 12 ans et plus dans celles d'Alario et Ferrand et de Bonin et collègues), et donc d'autoriser un travail avec des mots aux caractéristiques plus diverses sans perte de précision sur les âges plus élevés. Concernant la familiarité, cette dimension n'est renseignée que par les banques d'Alario et Ferrand, et de Bonin et collègues. Elle n'est donc actuellement disponible que pour 699 mots du Français. La méthode de recueil suivie ici diverge de celle suivie par les collègues sur trois points : l'utilisation d'une échelle catégorielle en 6 niveaux au lieu de 5 (ce qui permet d'éviter les réponses centrales de

facilité en cas de difficulté d'estimation, et ainsi d'offrir une précision plus poussée), l'évaluation de mots, et non d'images, et l'utilisation d'une consigne adaptée (pour les détails sur le choix de la consigne, se reporter à la section « Information recueillie et opérationnalisation du recueil »). En dépit du fait que CHACQFAM résulte de l'évaluation de mots, et non d'images, une précision optimale des indices a pu être obtenue grâce à la distinction des différentes acceptions sémantiques d'un mot lors de la présentation des items aux participants. Ainsi, l'information recueillie pour chaque item correspond bien au mot présenté et non à un autre, aussi bien dans les cas d'homographie que d'une mauvaise maîtrise de l'orthographe (confusions en cas d'homophonie). Par ailleurs, la technique de recueil utilisée a contribué à l'obtention d'une information de qualité. L'outil informatique spécifiquement développé (détail à la section « Méthode ») a en effet permis, d'une part d'échantillonner largement la population francophone sans limitation territoriale, et d'autre part d'éviter toute pression psychologique sur les participants (contribution volontaire et absence de contraintes de productivité), ce qui a certainement favorisé l'introspection nécessaire à l'estimation sérieuse de chaque item. Enfin, les données brutes ont fait l'objet d'un filtrage statistique, garantissant un ciblage optimal de la valeur centrale pour chaque item et indice.

Étant donné le choix réfléchi pour une procédure qui n'a pas totalement respecté le standard suivi par les travaux des collègues cités plus haut, eux-mêmes inspirés des travaux effectués dans le monde anglo-saxon, CHACQFAM est livrée avec une information détaillée sur la méthodologie suivie à toutes les étapes de son élaboration. Ceci permet à tout chercheur qui le souhaite, de mettre en œuvre les mêmes procédures pour constituer, valider ou compléter la banque existante.

La présentation de la banque de données CHACQFAM détaille dans l'ordre suivant les caractéristiques du corpus utilisé, l'information recueillie et l'opérationnalisation du

recueil, l'outil de recueil, l'enquête, les données brutes obtenues, les procédures de filtrage et de transformation des données brutes, les procédures de validation du recueil et du contenu de la base, et un examen quantitatif et qualitatif des données finales proposées dans la banque de donnée.

METHODE

Le corpus des 1225 mots utilisés

Le corpus employé comporte un ensemble originel de mots utilisés dans le cadre d'expériences de laboratoire sur la reconnaissance des mots parlés, correspondant à des mots du vocabulaire courant. Pour les caractéristiques détaillées de ces mots, ainsi que pour les raisons de leur sélection, le lecteur peut se référer au travail de Lachaud (Lachaud, 2005). S'ajoutent à ce corpus initial tous les mots enchâssés dans ces mots expérimentaux ainsi que les dupliquas formels des homophones de ces mots enchâssés, sélectionnés à partir du dictionnaire « Petit Larousse ». Le corpus comprend en majorité des noms, mais aussi d'autres catégories grammaticales (verbes conjugués, adjectifs, etc.). Un récapitulatif des propriétés générales formelles des mots de la base de données CHACQFAM est proposé dans le Tableau I.

INSERER TABLEAU I

Information recueillie et opérationnalisation du recueil

L'enquête a consisté à recueillir 3 types d'information de nature psychologique : une estimation de l'âge d'acquisition d'un mot, une estimation du degré de certitude sur l'estimation produite pour l'âge d'acquisition du mot, et une estimation de la familiarité du participant avec le mot présenté. Cette section décrit et justifie la manière dont ces 3 dimensions ont été opérationnalisées pour le recueil.

Âge d'acquisition estimé

La mesure de l'AoA a été effectuée à partir d'estimations par des adultes, au lieu d'une mesure directe auprès de populations de différentes tranches d'âge. Ce choix est justifié par quatre éléments : l'estimation par des adultes semble reflèter fidèlement l'AoA réel (voir plus bas), le protocole est moins lourd que pour une mesure d'AoA réel (population d'enfants difficile à obtenir et à enquêter, nécessité d'un échantillon très élevé dû à la démultiplication des tranches d'âge, etc.), le vocabulaire est plus varié que celui utilisable avec des enfants (intérêt supérieur pour des recherches menées ensuite avec des adultes), et c'est une procédure qui a été employée par plusieurs auteurs pour produire des normes d'AoA (voir par exemple Juhasz, 2005).

Étant donné qu'il existe un débat sur la validité d'une estimation de l'AoA par des adultes, par rapport à une mesure réelle (par exemple, Chalard et collègues (Chalard et al., 2003) montrent que l'âge d'acquisition objectif est un meilleur prédicteur des latences de production parlée et écrite que l'âge d'acquisition estimé), quelques éléments de réflexion sont proposés au lecteur pour justifier le choix de la procédure qui a été suivie.

Une estimation fournie par des adultes serait moins précise qu'une mesure directe pour deux raisons. La première raison est psychologique : imprécision de la mémoire, interférence sur l'estimation de divers paramètres comme la familiarité ou la fréquence d'usage, interférence de souvenirs reconstruits, etc. La deuxième raison est méthodologique. On peut par exemple relever, dans les études publiées, un manque de précision de l'échelle employée (tranches d'âge de deux, voire trois ans, au lieu d'un pas annuel ; catégorie supérieure unique pour tous les âges au-delà de la limite choisie), des conditions de passation ne permettant pas une introspection suffisante (en groupe, pression à la productivité dans un temps limité, quantité importante d'items à traiter en une fois, etc.), une absence de procédures de filtrage des données brutes, etc.

Par une correction de la méthodologie suivie, il est possible d'améliorer la précision de l'indice final, et un développement a eu lieu dans ce sens pour générer l'AoA de CHACQFAM. Par contre, les causes psychologiques à l'imprécision d'une estimation ne sont pas contrôlables. Cependant, un certain nombre d'éléments montre que les estimations produites par des adultes, même si elles sont moins précises que l'AoA objectif, reflètent avec une fidélité tout à fait acceptable l'âge d'acquisition réel d'un mot (Gilhooly et Gilhooly, 1980 ; Morrison et al., 1997). Par exemple, les effets obtenus avec l'âge d'acquisition objectif se retrouvent avec l'âge d'acquisition estimé, les deux facteurs étant fortement corrélés (r = .69, selon Chalard et collègues (Chalard et al., 2003), r = .76 selon Morrison et collègues (Morrison et al., 1997), r = .74 selon cette étude (corrélation avec les données d'âge d'acquisition réel de Chalard et al., 2003)). Une étude longitudinale (Jorm, 1991) a même pu démontrer la validité de l'âge d'acquisition estimé, en tant que reflet fidèle de l'âge d'acquisition objectif, tandis que les échantillons de population testés à travers différentes études présentent généralement une bonne consistance de leurs estimations (Morrison et al., 1997). De plus, l'AoA fourni par CHACQFAM se révèle être un très bon prédicteur des temps de reconnaissance de mots parlés isolés, une fois décorrélé de la fréquence objective et de la familiarité (Lachaud, soumis).

Aussi, l'AoA estimé est apparu comme un compromis convenable entre précision et lourdeur du protocole.

Dans le sondage réalisé pour la présente étude, il était demandé aux participants d'évaluer l'AoA selon la procédure habituellement employée pour recueillir ce type d'information (voir par exemple Ferrand et collègues. (Ferrand et al., 2003)).

L'attention des participants était tout particulièrement focalisée sur l'âge auquel ils pensaient avoir été mis en contact pour la première fois avec l'objet désigné par le mot présenté - objet physique ou concept (focalisation sur le signifié plutôt que sur le signifiant

des mots). L'emploi d'une telle procédure est imposé par le souhait d'optimiser la précision du rappel. Premièrement, lorsque l'individu découvre un objet, il cherche spontanément, dès l'enfance, à acquérir la forme lexicale qui le désigne, afin de pouvoir communiquer à son sujet avec d'autres humains. La corrélation dans le temps entre l'expérience de l'objet et l'acquisition du signifiant relatif doit donc être très élevée. Deuxièmement, le participant devrait avoir une plus grande facilité pour localiser dans le temps son vécu à un objet qu'à l'apprentissage d'une forme verbale pour le désigner. L'expérience concrète liée au signifié peut donc représenter une aide essentielle, et rendre la tâche plus facile. Enfin, focaliser les participants sur l'utilisation d'une stratégie particulière permet probablement une réduction de la variance dans les données. La standardisation adoptée ici se rapproche de plus de celle qui est implicitement imposée au sujet par une présentation d'image, puisque dans la plupart des normes d'âge d'acquisition, des images plutôt que des mots ont été employées pour générer les estimations (par exemple : Morrison et al., 1997 ; Alario et al., 1999 ; Bonin et al., 2003).

Degré de certitude de l'estimation pour l'âge d'acquisition

Le degré de certitude sur l'estimation de l'âge d'acquisition est un indice secondaire ajouté au recueil dans le but de disposer d'une information supplémentaire pour évaluer la qualité de l'estimation des âges d'acquisition. Il apporte donc une aide pour valider l'estimation, mais aussi pour sélectionner les mots dans la base de données lors de son utilisation (par exemple, il est possible de générer un gradient de qualité de l'estimation de l'AoA en utilisant de façon combinée l'écart-type de l'AoA et le degré de certitude).

Le degré de certitude de l'âge d'acquisition est simplement opérationnalisé en posant directement la question aux participants.

Familiarité

La familiarité d'un individu avec un mot dépend essentiellement de 2 aspects : la fréquence avec laquelle un mot ou le concept (ou l'objet) auquel il réfère est rencontré et

utilisé, et la complexité du concept (ou son degré d'abstraction) qu'il véhicule. Ces 2 aspects semblent pouvoir être évalués séparément grâce à la fréquence subjective pour le premier, et à la concrétude pour le second (par exemple Bonin et collègues (Bonin et al., 2003)).

Les normes de familiarité antérieurement publiées ont utilisé une consigne relativement standard à travers les études. Une traduction en français en est donnée ci-après (reprise de Barry et al., 1997) : « Jugez la familiarité de l'objet représenté sur l'image. Donnez un degré de familiarité selon le caractère habituel ou inhabituel de l'objet dans votre expérience quotidienne. Comprenez par familiarité le degré auquel vous entrez en contact avec, ou pensez à, l'objet représenté ». L'indice de familiarité tel qu'il est décrit dans cette consigne apparaît très voisin d'un indice de fréquence subjective. La mesure obtenue à partir de cette consigne ne représenterait donc pas correctement les diverses composantes de la familiarité. Toutefois, si les stimuli présentés aux participants sont des images, c'est que les concepts représentés ont un degré de concrétude plutôt élevé pour pouvoir être dessinés, et on peut donc supposer que la situation reste implicitement équilibrée entre la composante fréquentielle et la composante concrétude de la familiarité. Ce ne serait plus le cas avec un protocole présentant des mots, dont certains peuvent véhiculer des concepts abstraits et difficilement imageables, comme lors de l'enquête pour CHACQFAM. Une consigne qui focaliserait les participants sur les aspects fréquentiels de la familiarité conduirait alors probablement à un appauvrissement de l'indice final. De plus, étant donné que la fréquence subjective et la fréquence objective sont fortement corrélées (2), l'intérêt d'un recueil de la fréquence subjective est discutable parce qu'il risque de n'apporter qu'une quantité limitée d'information nouvelle, tandis que la concrétude d'un mot ne conviendrait pas puisqu'elle ne mesure pas directement la sensation de familiarité éprouvée par le sujet psychologique.

---

[2] $.41 \leq \underline{r} \leq .79$ d'après les études citées par Desrochers et Bergeron (Desrochers et Bergeron, 2000), impliquant que l'estimation de la fréquence subjective fournirait une information redondante jusqu'à près de 62% ($\underline{r}^2 = .62$) de celle déjà disponible dans les bases de données lexicales du français.

Le choix s'est donc porté sur une consigne demandant au participant d'exprimer directement sa familiarité avec le mot présenté à l'aide d'une échelle graduée. La consigne ne donne aucune autre spécification sur le terme « familiarité » que la connaissance qu'a le participant du mot (voir la consigne dans la section « Questionnaire d'enquête »), et ne précise pas le niveau de représentation sur lequel doit porter l'estimation (signifié ou signifiant). Par ce défaut de précision, il pourrait sembler que le participant se soit focalisé sur une estimation de sa familiarité pour le signifiant. Toutefois, ce n'est pas le cas pour deux raisons. D'une part, les items sont désambiguïsés par un synonyme, et donc la dimension sémantique est également présente. D'autre part, il est peu probable que signifiants et signifiés soient des unités indépendantes dans le lexique mental, une représentation lexicale étant au moins constituée de ces deux pôles, ce qui implique que si le signifié est activé, le signifiant l'est aussi, et inversement.

Cette consigne offre plusieurs avantages. Premièrement, elle n'est pas ambiguë, car tout un chacun peut y répondre naturellement et intuitivement. Deuxièmement, l'indice qu'elle permet de générer véhicule une information psychologique multidimensionnelle à même de servir directement de prédicteur de la difficulté du traitement d'un mot (l'impression d'être familier avec un mot est probablement basée à la fois sur des aspects fréquentiels, sémantiques, etc.). Cet indice peut aussi être considéré comme une variable dépendante, et faire l'objet d'une analyse statistique utilisant des indices objectifs pour en connaître les composantes, et pour en extraire l'influence de chaque composante sur la réponse fournie par les participants.

L'outil de recueil

Il s'agit d'une interface électronique structurée autour de 2 questionnaires (inscription et enquête), utilisable à l'aide d'un navigateur Internet. Cette interface était accessible à partir du site du Laboratoire de Psycholinguistique Expérimentale de l'Université de Genève durant

la période de recueil des données. Elle présente de l'information à l'utilisateur et récolte ses réponses. Ces dernières sont stockées au fur et à mesure qu'elles sont fournies, sur un serveur de l'Université de Genève, pour finalement constituer une base de données. Cette méthode de recueil, moderne et efficace, a permis d'échantillonner largement la population francophone en France et en Suisse Romande, et d'obtenir rapidement un corpus important et de qualité. Le système a été conçu pour obtenir des informations générales sur le participant, utilisées pour la validation des données brutes lors de leur traitement. Il a également été conçu pour être souple pour les utilisateurs. La quantité d'items à traiter étant importante, il était peu probable qu'une même personne les traite tous. Il était encore moins probable qu'elle les traite tous en une seule session de travail si elle choisissait de faire l'enquête dans sa totalité. Il fallait donc offrir la possibilité à quelqu'un ne disposant pas de beaucoup de temps, mais volontaire, de pouvoir s'impliquer de manière cumulée en reprenant à n'importe quel moment une enquête interrompue. C'est pour cette raison qu'un système d'identification anonyme a été développé. Enfin, le système devait autoriser l'analyse statistique des données récoltées. Les items ont donc été présentés aléatoirement à chaque participant, une seule fois chacun, que l'enquêté ait procédé en une ou plusieurs sessions de travail. Théoriquement, chaque item est en moyenne traité un nombre équivalent de fois lorsque le nombre de participants est élevé. De plus, les données sont déjà sous un format électronique, et ne nécessitent pas une saisie fastidieuse avant leur traitement.

Le questionnaire d'inscription

Le participant qui commence l'enquête doit d'abord créer son identité anonyme au sein du système. Celle-ci résulte de l'association entre un profil personnel et une clé d'accès (identifiant numérique). Le profil personnel comprend des informations sur la tranche d'âge, le sexe, le niveau d'éducation, l'origine linguistique et la compétence linguistique du participant (listes de choix à renseigner obligatoirement). Ces informations sont absolument

anonymes, et ne permettent en aucun cas, directement ou par recoupement, de connaître la source d'émission des données, qu'il s'agisse d'individus ou d'organisations. Elles sont utilisées pour filtrer les données brutes (voir plus loin, « Procédures de transformation des données brutes »). L'identifiant numérique correspond à une combinaison classique de type login + mot de passe, tous deux choisis par l'utilisateur. À cette identité anonyme est associée la liste des items restant à traiter spécifiquement par ce participant (soit, lors de la première connexion, les 1225 mots du corpus).

Le système présente une faille potentielle. En effet, en cas de perte de la clé numérique personnelle (login, mot de passe, ou combinaison des deux), un participant qui souhaite poursuivre l'enquête doit à nouveau remplir le questionnaire d'inscription, puis créer une nouvelle identité dans le système (ce dernier refuse tout doublon de clé numérique). Dans ce cas de figure, des items déjà traités sous l'ancienne identité peuvent de nouveau être présentés au même individu, engendrant l'enregistrement de plusieurs réponses de ce participant à un même item. Toutefois, il était très peu probable que ce biais survienne. Premièrement, il a été explicitement demandé aux participants, dans la page de présentation de l'étude, d'arrêter l'enquête s'ils se trouvaient dans ce cas de figure. Deuxièmement, l'enquête elle-même n'est pas un exercice des plus attrayants. Il est donc psychologiquement peu vraisemblable, compte tenu de l'effort nécessaire, qu'un participant remplisse à nouveau le questionnaire d'inscription. Troisièmement, les participants savaient qu'ils contribuaient à une étude scientifique, de laquelle ils n'allaient retirer aucune autre satisfaction que celle d'avoir aidé au développement de la Connaissance. Ils savaient également qu'ils auraient accès aux résultats de l'enquête dès que les conclusions seraient disponibles (via le site web utilisé pour l'enquête), ce qui pouvait représenter une motivation. Quatrièmement, la quantité de mots à traiter étant importante (N = 1225) et le pourcentage moyen de mots traités par participant étant faible (N = 142), la probabilité d'avoir recueilli des doublons est encore restreinte.

Enfin, vu la grande qualité des données obtenues, il est plutôt probable que les participants aient travaillé très sérieusement.

Le questionnaire d'enquête

Le questionnaire d'enquête présente au participant un mot écrit dont le sens est précisé entre parenthèses par un synonyme (exemple : « air (mélodie) »). Le participant utilise ce questionnaire pour fournir l'âge auquel il pense avoir appris ce mot et le degré de certitude de son estimation puis son degré de familiarité avec l'item. L'âge d'acquisition est donné librement (choix ouvert, réponse non optionnelle), avec un format numérique en années (nombre entier à 1 ou 2 chiffres), sans possibilité de répondre « ne sais pas ». La consigne écrite était : « Vers quel âge pensez-vous avoir appris ce mot ? Estimez l'âge auquel vous avez appris ce mot (âge auquel vous pensez avoir été mis pour la première fois en contact avec l'objet réel ou le concept correspondant à l'item) ». Le degré de certitude sur la réponse fournie pour l'âge d'acquisition estimé est recueilli à l'aide d'une liste de 6 choix (cases à cocher), du moins certain (---) au plus certain (+++). La réponse est non optionnelle. La consigne écrite était : « Cochez la case correspondant à votre degré de certitude sur l'âge évalué », et le participant voyait sous la série de cases la légende suivante : « (moins certain) --- / -- / - / + / ++ / +++ (plus certain) ». Le degré de familiarité avec le mot est recueilli selon la même procédure que celle utilisée pour recueillir le degré de certitude (réponse non optionnelle dans une liste de 6 choix sur une échelle du même type que précédemment – consigne écrite : « Quel est votre degré de familiarité avec ce mot ? Estimez le degré de familiarité que vous avez avec ce mot (est-ce un mot que vous connaissez parfaitement ou est-ce un mot qui vous est inconnu ?) »).

Lorsqu'un item a été traité sur les 3 dimensions, le participant valide ses réponses en cliquant sur un bouton. Cette action a pour effet d'enregistrer les informations dans la base de données et de présenter l'item suivant. Le questionnaire présente aléatoirement les items un

par un. La liste résiduelle (liste de laquelle l'item présenté a été tiré moins l'item qui vient d'être présenté) est ainsi parcourue, jusqu'à ce qu'elle soit vide.

L'enquête

L'enquête a été lancée à l'aide d'une campagne de communication incitative auprès de 53 universités de France métropolitaine et de Suisse romande, dont seulement une partie a répondu positivement en diffusant l'information auprès de son personnel et de ses étudiants. Elle n'a pas duré plus de 3 mois, ayant pris fin en juillet lorsque les participants ont cessé d'avoir une activité sur le site.

PRESENTATION DES DONNEES BRUTES ET DES PROCEDURES DE TRAITEMENT

Présentation des données brutes

Les données brutes ont été générées par 196 participants (121 femmes et 75 hommes), majoritairement âgés de 18 à 25 ans (N = 97 - répartition pour les autres tranches d'âge : 3 non majeurs, 50 âgés de 25 à 35 ans, 38 de 35 à 55 ans, et 8 de plus de 66 ans). Au total, 27778 réponses ont été recueillies. Chaque mot a été traité en moyenne par 22,68 participants. Le participant moyen a traité environ 142 mots.

Procédures de transformation des données brutes

Filtrage des participants

Dans un premier temps, les participants ont été filtrés sur la base de 2 critères. Le premier critère est linguistique. Ce filtre utilise les profils linguistiques renseignés à l'aide du premier questionnaire de l'enquête, pour ne conserver que les individus francophones unilingues, ou bilingues avec un excellent niveau de maîtrise de la langue française. Le deuxième critère est un critère statistique de normalité des réponses par mot. Ce filtre permet de détecter les individus « aberrants », c'est-à-dire ceux ayant un nombre de réponses

excentrées par rapport à la distribution des réponses des autres participants sur un nombre élevé d'items ou sur des items qui ne doivent pas causer de difficulté pour la tâche. En effet, le premier filtre pouvait laisser passer un certain nombre de cas (exemple : individus se caractérisant comme francophones unilingues d'après les informations renseignées par le questionnaire d'inscription, mais ayant une maîtrise douteuse de la langue française. On trouve ainsi dans leurs réponses des mots très courants comme « œufs » estimés appris à 34 ans). L'application de ces 2 critères a conduit à éliminer 26 participants, à la fois non francophones (8 de langue maternelle allemande, 1 de langue maternelle anglaise, 4 de langue maternelle espagnole, 5 de langue maternelle italienne, et 8 de langue maternelle non répertoriée dans la liste à choix), et présentant des distributions curieuses. Les 170 participants restant sont 151 français unilingues, 11 bilingues par le milieu social (la langue maternelle est le Français, la langue seconde est apprise dès après 7 ans sous l'influence du contexte social ou scolaire : Allemand, N = 5 ; Anglais N = 4 ; autre, N = 2), et 8 bilingues par le milieu familial (les deux langues sont apprises dès avant 4 ans : Français / [Anglais, N = 2 ; Espagnol, N = 1 ; Italien, N = 1 ; Portugais, N = 2 ; Autre, N = 2]).

Filtrage par item des réponses des 170 participants restants

La distribution des données a été vérifiée pour chaque item de manière à filtrer toute réponse anormale. Quelques réponses ont encore été éliminées (exemples : réponses 0 an ou 99 ans, probablement pour « ne sais pas », réponses non normales et suspectes, comme « trac (anxiété) » appris à 2 ans (un enfant de 2 ans peut avoir peur, mais peu vraisemblablement avoir le trac), etc.). Au total, 2867 réponses ne satisfaisant pas aux critères statistiques et qualitatifs définis plus haut ont été éliminées. La base de données présentée ici a donc été

constituée à partir de 24911 réponses (de 11 à 32 réponses par item ; moyenne = 20.34 ; écart-type = 3.73) (3).

Notons que, si le nombre moyen de réponses par item est plus faible dans CHACQFAM (N = 20) que dans les bases d'Alario et collègue (N = 26 pour l'âge d'acquisition, N = 30 pour la familiarité) ou de Ferrand et collègues (N = 40), la taille de l'échantillon reste acceptable si on s'en réfère à la base de Bonin et collègues, qui présente une information d'âge d'acquisition estimée à partir d'un échantillon de 20 personnes. Par ailleurs, la perte éventuelle de précision de l'estimation à cause d'un échantillon plus faible est compensée par une plus grande précision des échelles utilisées dans CHACQFAM que dans les banques de données citées en référence (respectivement, pas annuel pour l'estimation de l'âge d'acquisition et échelle en 6 points pour l'estimation de la familiarité, pas bi ou trisannuel pour l'estimation de l'âge d'acquisition et échelle en 5 points pour l'estimation de la familiarité). Elle est également compensée par l'existence d'une procédure de filtrage des données brutes. De plus, pour les items ayant un échantillon faible (N < 15), l'examen détaillé des données ne permet pas de relever d'écart moyen remarquable par rapport aux mêmes items des bases existantes (écart avec la base de Bonin et collègues : N = 10 pour le nombre de mots communs entre bases, différence moyenne de 0.001 pour l'indice de familiarité, de 0.124 pour l'âge d'acquisition estimé ; écart avec la base de Ferrand et collègues : N = 8, différence moyenne de 0.937 pour l'âge d'acquisition estimé ; écart avec la base de Alario et collègue : N = 5, différence moyenne de 0.243 pour l'indice de familiarité, de 0.548 pour l'âge d'acquisition estimé). Enfin, il faut préciser que la quantité d'items pour lesquels l'échantillon est faible (inférieur ou égal à 15) est peu importante (11 réponses par mot : 4 items (0.33%) ; 12 réponses : 7 items (0.57%) ; 13 réponses : 20 items (1.6%) ; 14

---

[3] Caractéristiques de la distribution brute avant filtrage : 27778 réponses, 12 à 34 réponses par item, moyenne = 22.68, écart-type = 3.89.

réponses : 33 items (2.7%) ; 15 items : 51 items (4.1%) – soit au total 115 items ou 9.4% de la base). Par conséquent, ces items ont été laissés dans la base, tout en permettant aux utilisateurs de CHACQFAM qui souhaiteraient ne pas les utiliser, de disposer de toute l'information nécessaire pour les identifier.

Codage des données filtrées et extraction de l'information

Les réponses fournies pour la certitude de l'âge d'acquisition estimé et pour la familiarité ne sont pas numériques, contrairement aux réponses données pour l'âge d'acquisition. Elles ont donc été codées numériquement pour permettre de réaliser les calculs de moyenne et d'écart type (--- = 1 ; -- = 2 ; - = 3 ; + = 4 ; ++ = 5 ; +++ = 6). Le nombre de réponses par item (N) a été recensé. Pour chaque item et pour chaque dimension psychologique étudiée (âge d'acquisition estimé, degré de certitude pour l'âge d'acquisition fourni, familiarité), la moyenne (M), la valeur minimum (Min), la valeur maximum (Max) et l'écart type (E.T.) ont été extraits. Ces indices forment le corps de la base de données pour chacun des facteurs (M, Min, Max, E.T., en plus de N).

VALIDATION DE L'INFORMATION CONTENUE DANS LA BASE DE DONNEES CHACQFAM

L'information contenue dans la banque de données CHACQFAM a subit une double procédure de validation. La première concerne la vérification de la validité du recueil, la deuxième concerne la vérification de la conformité des données avec les normes existantes. Le ciblage de la valeur centrale a de plus été examiné.

Vérification de la validité du recueil

Dans le cas où le nombre de réponses fournies pour chaque mot varierait en fonction de l'âge d'acquisition estimé, de la familiarité, ou du degré de certitude sur l'âge estimé, c'est qu'un biais pourrait exister dans la production des réponses. En effet, le système de recueil

ayant été testé avant sa mise en production pour s'assurer de la présentation aléatoire des items, le nombre de réponses par item ne doit pas varier en fonction de l'un des 3 facteurs. Si des variations sont observées, c'est que le comportement des participants est influencé par l'une de ces 3 dimensions. Pour s'assurer que ce n'est pas le cas, une analyse statistique de type analyse de régression linéaire a été effectuée (variable dépendante = nombre de réponses recueillies par item ; prédicteurs = âge d'acquisition estimé, degré de certitude de l'estimation fournie pour l'âge d'acquisition, familiarité, interactions 2 à 2 entre ces 3 prédicteurs, interaction triple, et covariance entre prédicteurs). Les résultats du test de significativité pour les effets fixes sont présentés dans le Tableau II.

INSÉRER TABLEAU II

L'analyse révèle que la quantité de réponses par item varie en fonction du degré de certitude de l'estimation pour l'âge d'acquisition. Plus le degré de certitude augmente et plus la quantité de réponses par item augmente (augmentation moyenne de 1.1 réponses par degré de certitude supplémentaire ; effet significatif à un seuil $p < .0001$). Aucun autre facteur n'a d'influence significative, aucune interaction n'est significative.

Il existe donc un biais quantitatif dans les données, puisque le nombre de réponses fournies par les participants pour chaque item dépend de la difficulté qu'ils ont eue à évaluer l'âge auquel ils ont appris le mot. Ce phénomène peut s'expliquer par le choix des participants de se déconnecter préférentiellement au moment où la difficulté pour répondre devient plus grande, ce facteur ayant pu être renforcé par un degré de lassitude élevé face à une tâche répétitive. Toutefois, il faut mettre en avant le fait qu'aucune relation n'apparaît entre la quantité de réponses recueillie et les 2 facteurs d'intérêt principal, l'âge d'acquisition estimé et la familiarité. Les données recueillies pour ces 2 facteurs apparaissent donc validées par cette première procédure de vérification.

Vérification de la conformité des données avec les normes existantes

La deuxième procédure de validation est une comparaison des données de CHACQFAM avec chacune des 4 bases prises comme référence (Alario et collègue (Alario et Ferrand, 1999), Ferrand et collègues (Ferrand et al., 2003), Bonin et collègues (Bonin et al., 2003), Chalard et collègues (Chalard et al., 2003)). Rappelons en effet que CHAQFAM a un certain nombre de mots en commun avec ces 4 banques de données. Les résultats de cette comparaison, obtenus par une analyse de régression linéaire, sont récapitulés dans le Tableau III. Ils montrent généralement une forte corrélation (R^2 ajusté de .58 à .70 pour l'âge d'acquisition estimé, de .54 pour une comparaison avec l'âge d'acquisition réel, et de .07 à .43 pour la familiarité).

INSÉRER TABLEAU III

Parallèlement, les 4 bases sont comparées entre elles grâce à des analyses de régression linéaire simple (Tableau IV – R^2 ajusté de .91 à .94 pour l'âge d'acquisition estimé, et de .11 à .29 pour une comparaison entre âge d'acquisition estimé et réel. Pour la comparaison des scores de familiarité, seules les bases d'Alario et collègue et de Bonin et collègues pouvaient être utilisées, mais d'une part le nombre d'items communs entre les deux bases était trop faible (N = 6), et d'autre part seulement 2 items sur 6 (« léopard » et « tonneau ») avaient une information de familiarité dans la base d'Alario et collègue (Alario et collègue vs Bonin et collègues : respectivement 2.45 vs. 1.37, et 2.4 vs. 1.27)).

INSÉRER TABLEAU IV

Enfin, une comparaison est effectuée entre les résultats des analyses de régression impliquant CHACQFAM et les résultats des analyses de régression n'impliquant pas CHACQFAM. Elle montre que les données d'âge d'acquisition estimé contenues dans la base CHACQFAM ont des caractéristiques intermédiaires entre les données d'âge d'acquisition estimé par Alario et collègue, Ferrand et collègues, ou Bonin et collègues (avec lesquelles elles corrèlent moins bien que les données de ces bases de référence entre elles), et les

données d'âge réel mesurées par Chalard et collègues (avec lesquelles elles corrèlent mieux que les données des bases de référence précédentes ne le font). Ce résultat indique que l'estimation de l'âge d'acquisition fournie par CHACQFAM est plus proche de l'âge réel d'acquisition que les estimations fournies par les bases d'Alario et collègue et Ferrand et collègues (pas de statistiques pour Bonin et collègues). Concernant les écarts entre CHACQFAM et les bases de référence pour la familiarité, l'explication qui peut être avancée concerne les différences de procédure de recueil entre études (échelles, tri des données brutes, conditions de passation, précision des acceptions sémantiques, etc.). Hélas, le nombre de mots en commun entre les bases d'Alario et collègue et Bonin et collègues est trop faible pour faire une analyse statistique. On remarque toutefois un écart notable entre les 2 items en commun bien que la procédure de recueil soit similaire dans ces 2 études (mais : seulement 2 items).

Ciblage de la moyenne

L'examen du ciblage de la moyenne fournit une indication de la congruence des réponses des participants. Concernant l'estimation de l'âge d'acquisition, l'écart-type moyen est de 2.7 (Min = 0.9, Max = 10.9, E.T. = 1.2). Il augmente avec l'âge d'acquisition estimé (β = .78). Concernant le degré de familiarité, il est de 1.1 (Min = 0, Max = 2.9, E.T. = 0.6). Il diminue lorsque le degré de certitude augmente (β = -.85).

Cette information permet de confirmer une congruence plutôt bonne des réponses des participants, tant pour l'estimation de l'âge d'acquisition que pour le degré de Familiarité.

Conclusions sur la validité des données de la base CHACQFAM

L'information proposée par CHACQFAM apparaît fiable. Premièrement, aucun biais au niveau des procédures de recueil d'âge d'acquisition ou de familiarité n'a pu être mis en évidence. Deuxièmement, l'âge d'acquisition estimé dans CHACQFAM reflète plus fortement l'âge réel d'acquisition mesuré par Chalard et collègues (Chalard et al., 2003) que de l'estimation trouvée dans les autres bases de référence. Enfin, ce facteur se révèle être très

fortement explicatif des temps de reconnaissance de mots isolés une fois décorrélé de la familiarité et de la fréquence objective (Lachaud, soumis).

Concernant la familiarité, il est a priori plus difficile de se prononcer, étant donné qu'il manque une référence à laquelle comparer le résultat des corrélations de CHACQFAM avec ces 2 bases. Cependant, l'indice Fam de la banque de données CHACQFAM s'avère à même de prédire les temps de reconnaissance de mots isolés avec un seuil de significativité voisin de 0, une fois l'AoA estimé et la fréquence objective décorrélées (Lachaud, soumis).

Considérant les résultats sur la validité du recueil et sur la validité de l'âge d'acquisition, sur la congruence des réponses pour l'AoA et pour la Fam (ciblage de la valeur moyenne), et sur le degré de significativité très élevé des deux indices de CHACQFAM pour prédire les temps de reconnaissance des mots, la banque de données CHACQFAM se présente comme un outil de qualité importante pour la recherche.

PRESENTATION DES DONNEES CONTENUES DANS LA BASE CHACQFAM

Examen quantitatif : quelques distributions

Âge d'acquisition estimé

INSÉRER FIGURE 1

En majorité, le vocabulaire testé est estimé appris entre 4 et 12 ans (moyenne : 8.74 ans ; écart-type : 3.47 ans) (4). Il semble, d'après cette distribution, que vers 16-17 ans, le lexique mental d'un individu de langue maternelle française soit à peu près en place (chute brutale de la quantité estimée de nouveaux mots appris vers 12-13 ans, et atteinte d'un pallier vers 16-17 ans), pour des mots courants à relativement courants..

---

[4] Les adultes rapportent peu de souvenirs antérieurs à l'âge de 3-4 ans (amnésie infantile). Il peut donc exister une distorsion des âges estimés aux alentours de la borne inférieure, qui serait masquée dans une norme basée sur un recueil catégoriel avec des tranches d'âge de 3 ans. Toutefois, on trouve dans CHACQFAM beaucoup de mots pour lesquels la valeur minimale d'âge estimé est 1 ou 2 ans.

Degré de certitude de l'estimation de l'âge d'acquisition

Les participants se sont sentis plutôt certains qu'incertains de leur évaluation des âges d'acquisition (moyenne positive : 0.34 [/3] ; écart-type : 0.49). La dispersion autour de la moyenne, symétrique, indique cependant qu'il existe aussi bien des réponses estimées avec une forte certitude que des réponses estimées avec une incertitude assez marquée.

Familiarité des participants avec les mots testés

Les participants se sont sentis familiers à très familiers avec la majorité des mots testés (moyenne : 2.34 [/3] ; écart-type : 0.67).

Examen qualitatif des réponses

Âge d'acquisition estimé

Les valeurs obtenues pour l'âge d'acquisition estimé, la certitude sur l'âge d'acquisition estimé, et la familiarité, présentent une grande plausibilité psychologique, ainsi qu'une finesse certaine. Concernant l'âge d'acquisition, ces qualités se distinguent plus aisément dans le cas d'homophones, pour lesquels des différences pertinentes se remarquent en fonction du sens. Par exemple : [« air » (mélodie) : 5.63 ans / « air » (gaz) : 6.07 / « aire » (surface) : 9.12 / « ère » (période) : 9.91 / « erre » (errer) : 10.39] ; [« balle » (ballon) : 3.3 ans / « balle » (munition) : 6.18 / « balle » (ballot) : 9.67 / « balle » (monnaie) : 10.71] ; [« ton » (tien) : 3.94 ans / « thon » (poisson) : 6.62 / « ton » (tonalité) : 9.62 / « ton » (coloris) : 9.76] ; [« pain » (aliment) : 4.13 ans / « pin » (végétal) : 7.25 / « pain » (coup) : 9.3 / « pain » (masse) : 11.72] ; [« loup » (mammifère) : 4.5 ans / « loup » (masque) : 10.09 / « loup » (poisson) : 12.32] ; [« porc » (animal) : 5.77 ans / « port » (maritime) : 6.21 / « port » (porter) : 10.56 / « porc » (débauché) : 11.11 / « port » (silhouette) : 12.78 / « pore » (orifice) : 13]. Les informations se révèlent être très intéressantes d'un point de vue développemental, tant sur le plan de l'acquisition du lexique, que sur la maturation de la personnalité et l'évolution du vécu de l'enfant au cours de son histoire (accès aux concepts, centres d'intérêt,

etc.). Un examen de la distribution des mots pour l'âge moyen d'acquisition estimé révèle que les premiers mots acquis, de 2.8 à 4 ans, sont majoritairement monosyllabiques, et véhiculent des concepts simples ou élémentaires. Soit ils sont directement liés à l'enfant à travers son anatomie (« dent », « œil », « doigt », « tête », « pied », etc.), la notion d'individualité (« moi », « toi », « tu », « il », « ma », « ton », « est », « suis », etc.), les caractéristiques du corps (« grand », « long », « gros », « âge », etc.), le comportement (« sage », « cri », « rire », etc.), le ressenti (« aïe » / « ouille » (douleur), « froid », « aime », etc.), soit ils sont en rapport étroit avec l'environnement dans lequel l'enfant évolue, comme les objets qu'il manipule (« eau », « lait », « balle », « jouet », « lit », etc.), les phénomènes qui surviennent dans cet environnement (« nuit », « noir », « jour », « neige », « pluie », « vent », « bruit », etc.), les êtres vivants qu'il y rencontre (« homme », « femme », « père », « chat », « fleur », « arbre », etc.). À l'opposé, les mots estimés acquis le plus tardivement, de 17 à 24 ans, sont monosyllabiques ou bisyllabiques (cette étude ne teste pas de mots de plus de 2 syllabes), ne sont pas forcément plus complexes au niveau formel, mais surtout, font référence à des éléments caractérisés par leur abstraction. Celle-ci peut être liée à l'éloignement des objets/êtres/phénomènes désignés, hors du quotidien de l'individu (« roussette », « phénol », « calice », « anche », « attique », « palette », « serveur », « Brig » (ville), « crack » (drogue), etc.) comme au degré d'élaboration des concepts que véhiculent ces mots (« lemme », « tore », « orbe », « tripe » (cigare), « ose » (glucide), etc.).

Degré de certitude de l'estimation de l'âge d'acquisition

Concernant le score de certitude sur l'estimation de l'âge d'acquisition, l'examen de la distribution des réponses moyennes par item pour les 1225 mots, en fonction de l'âge d'acquisition estimé et du degré de certitude, montre que la période de la vie où les participants pensent avoir appris la plus grande quantité des mots présentés (environ 4 à 12 ans) est aussi celle pour laquelle le degré de certitude est le plus bas. Ce phénomène peut être

lié à l'augmentation importante de la quantité de mots nouvellement mémorisés durant cette période, ce qui les rend plus difficiles à caractériser individuellement et à localiser dans le temps. Il peut aussi être lié à un phénomène développemental, un lexique mental en phase d'élaboration subissant peut-être, au gré de l'intégration de nouveaux éléments, un remaniement interne aboutissant à une mémoire plus confuse.

Malgré ce fait, le degré de certitude de l'estimation reste théoriquement utilisable pour évaluer la netteté du souvenir, ou du souvenir imaginé, logique, plausible, reconstruit, etc., ainsi que la cohérence des répondants. Par exemple, d'après cet indice, l'âge d'acquisition estimé pour le mot « pain » (masse) serait moins certain que celui estimé pour le mot « jeu » (divertissement), puisque le premier est marqué d'un score de certitude moyen plutôt faible (-0.94) et dispersé (de –3 à 3 ; E.T. : 1.7), le second d'un score de certitude moyen plutôt élevé (1.63) et plus homogène (de –1 à 3, E.T. : 1.3). Parallèlement, l'âge d'acquisition moyen est peu ciblé pour le mot « pain » (masse) (minimum : 4 ans ; maximum : 26 ans ; E.T. : 6.4), alors qu'il apparaît plutôt ciblé pour le mot « jeu » (divertissement) (de 1 à 6 ans, E.T. : 1.3).

Familiarité des participants avec les mots testés

Concernant la familiarité, il semble exister un lien très fort entre ce facteur et l'âge d'acquisition estimé, visible sans autre analyse : les mots non familiers tendent tous à avoir des valeurs élevées d'âge d'acquisition estimé (r = -.78). L'observation qualitative des données suggère donc qu'un mot appris très tôt dans l'enfance, du fait de sa référence à des éléments conceptuellement simples et de sa grande antériorité (on est plus familier avec des objets que l'on connaît depuis longtemps), devrait être considéré comme très familier. Ceci se vérifie mathématiquement à partir des scores de la base de données (pour les 57 mots estimés appris avant 4 ans, le degré de familiarité moyen est de 2.92).

Conclusions sur la méthode de recueil et sur la qualité de l'information obtenue

L'examen quantitatif des données confirme les avantages de la méthode de recueil adoptée dans cette recherche. Face à une information difficile à obtenir, la méthode s'avère efficace (échantillonnage de la population sur un vaste territoire de la francophonie) et économique (recueil rapide d'une quantité importante d'information avec un coût nul en terme d'expérimentation directe et de saisie des données).

L'examen qualitatif des données montre que l'avantage quantitatif n'a pas lieu au détriment de la qualité. La méthode est précise (l'information recueillie est très fine, comme le montre la comparaison des scores pour les homophones) et paraît fiable (l'information recueillie est cohérente, et consistante avec celle proposée par d'autres études).

CONCLUSIONS DE L'ETUDE

Ce travail a permis de proposer une méthode d'enquête pour constituer rapidement une base de données lexicales volumineuse pour deux facteurs psychologiques difficiles à recueillir, l'âge d'acquisition estimé et la familiarité avec un mot. Il permet d'offrir à la communauté psycholinguistique une base de données de 1225 mots monosyllabiques et bisyllabiques du Français (CHACQFAM), renseignant ces deux facteurs. La base de données CHACQFAM, gratuitement à disposition, peut être téléchargée à partir de l'adresse Internet « http://psycholinguistique.unige.ch/ ».

BIBLIOGRAPHIE

Alario F. X., Ferrand L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. Behavior Research Methods, Instruments and Computers, 31(3), 531-552.

Alario F. X., Ferrand L., Laganaro M., New B., Frauenfelder U. H., Segui J. (2004). Predictors of picture naming speed. Behavior Research Methods, Instruments and Computers, 36(1), 140-155

Barry C., Gerhand S. (2003). Both concreteness and age-of-acquisition affect reading accuracy but only concreteness affects comprehension in a deep dyslexic patient. Brain and Language, 84(1), 84-104.

Barry C., Hirsh K., Johnston R., Williams C. (2001). Age of acquisition, word frequency, and the locus of repetition priming of picture naming. Journal of memory and language, 44(3), 350-375.

Barry C., Morrison C. M., Ellis A. W. (1997). Naming the snodgrass and vanderwart pictures: Effects of age of acquisition, frequency and name agreement. Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 50A(3), 560-585.

Baumeister A. A. (1985). Age of acquisition and meaningfulness as predictors of word availability. Journal of General Psychology, 112(1), 109-112.

Bird H., Franklin S., Howard D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. Behavior Research Methods, Instruments and Computers, 33(1), 73-79.

Bonin P., Chalard M., Meot A., Fayol M. (2002). The determinants of spoken and written picture naming latencies. British Journal of Psychology, 93(Pt 1), 89-114.

Bonin P., Fayol M., Chalard M. (2001). Age of acquisition and word frequency in written picture naming. The Quarterly Journal of Experimental Psychology A, 54(2), 469-489.

Bonin P., Méot A., Aubert L., Malardier N., Niedenthal N., Capelle-Toczek M.-C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots. L'année Psychologique, 104, 655-694.

Bonin P., Peereman R., Malardier N., Méot A., Chalard M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. Behavior Research Methods, Instruments and Computers, 35(1), 158-167.

Brown G. D., Watson F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. Memory and Cognition, 15(3), 208-216.

Brysbaert M. (1996). Word frequency affects naming latency in Dutch when age of acquisition is controlled. European Journal of Cognitive Psychology, 8(2), 185-193.

Brysbaert M., Lange M., Van Wijnendaele I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. European Journal of Cognitive Psychology, 12(1), 65-85.

Brysbaert M., Van Wijnendaele I., De Deyne S. (2000). Age-of-acquisition effects in semantic processing tasks. Acta Psychoogica, 104(2), 215-226.

Chalard M., Bonin P., Méot A., Boyer B., Fayol M. (2003). Objective age-of-acquisition (aoa) norms for a set of 230 object names in French: Relationships with psycholinguistic variables, the English data from Morrison et al. (1997), and naming latencies. European Journal of Cognitive Psychology, 15(2), 209-245.

Connine C. M., Mullennix J., Shernoff E., Yelen J. (1990). Word familiarity and frequency in visual and auditory word recognition. Journal of Experimental Psychology: Learning, Memory and Cognition, 16(6), 1084-1096.

Content A., Mousty P., Radeau M. (1990). BRULEX. Une base de données lexicales informatisée pour le Français écrit et parlé. L'année Psychologique, 90(4), 551-566.

Desrochers A., Bergeron M. (2000). Valeurs de fréquence subjective et d'imagerie pour un échantillon de 1916 substantifs de la langue française. Revue Canadienne de Psychologie Expérimentale, 54(4), 274-325.

Ellis A. W., Morrison C. M. (1998). Real age-of-acquisition effects in lexical retrieval. Journal of Experimental Psychology: Learning, Memory and Cognition, 24(2), 515-523.

Ferrand L., Grainger J., New B. (2003). Normes d'âge d'acquisition pour 400 mots monosyllabiques. L'année Psychologique, 104, 445-468.

Garlock V. M., Walley A. C. (2001). Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. Journal of memory and language, 45, 468-492.

Gerhand S., Barry C. (1998). Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise. Journal of experimental psychology: learning, memory and cognition, 24(2), 267-283.

Gerhand S., Barry C. (1999 a). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. Memory and Cognition, 27(4), 592-602.

Gerhand S., Barry C. (1999 b). Age-of-acquisition and frequency effects in speeded word naming. Cognition, 73(2), B27-36.

Gilhooly K. J., Gilhooly M. L. (1980). The validity of age-of-acquisition ratings. British Journal of Psychology, 71(1), 105-110.

Hirsh K., Funnell E. (1995). Those old, familiar things: Age of acquisition, familiarity and lexical access in progressive aphasia. Journal of neurolinguistics, 9(1), 23-32.

Izura C., Ellis A. W. (2002). Age of acquisition effects in word recognition and production in first and second languages. Psicologica, 23(2), 245-281.

Jorm A. F. (1991). The validity of word age-of-acquisition ratings: A longitudinal study of a child's word knowledge. British journal of developmental psychology, 9(3), 437-444.

Juhasz B. J. (2005). Age-of-acquisition effects in word and picture identification. Psychological bulletin, 131(5), 684-712.

Kremin H., Hamerel M., Dordain M., De Wilde M., Perrier D. (2000). Age of acquisition and name agreement as predictors of mean response latencies in picture naming of French adults. Brain and cognition, 43(1-3), 286-291.

Lachaud C. M. (2005). La prégnance perceptive des mots parlés : Une réponse au problème de la segmentation lexicale ? Thèse de Doctorat, Genève, GE, Université de Genève, Suisse. http://www.unige.ch/ cyberdocuments/theses2005/LachaudC/meta.html

Lachaud C. M. (soumis). Statistical investigation of the relation between objective word frequency, estimated age of word acquisition, familiarity with a word, and lexical recognition latency, in the processing of isolated French spoken words by normal adults. Cognitive Psychology.

Lyons A. W., Teer P., Rubenstein H. (1978). Age-at-acquisition and word recognition. Journal of psychological research, 7(3), 179-187.

Meschyan G., Hernandez A. (2002). Age of acquisition and word frequency: Determinants of object-naming speed and accuracy. Memory and cognition, 30(2), 262-269.

Monaghan J., Ellis A. W. (2002). Age of acquisition and the completeness of phonological representations. Reading and writing: an interdisciplinary journal, 15, 759-788.

Morrison C. M., Chappell T., Ellis A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 50A(3), 528-559.

Morrison C. M., Ellis A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. Journal of experimental psychology: learning, memory and cognition, 21(1), 116-133.

Morrison C. M., Ellis A. W. (2000). Real age of acquisition effects in word naming and lexical decision. British Journal of Psychology, 91(2), 167-180.

Morrison C. M., Hirsh K., Chappell T., Ellis A. W. (2002). Age and age of acquisition: An evaluation of the cumulative frequency hypothesis. European Journal of Cognitive Psychology, 14(4), 435-459.

Morrison C. M., Hirsh K. W., Duggan G. B. (2003). Age of acquisition, ageing, and verb production: normative and experimental data. The Quarterly Journal of Experimental Psychology A, 56(4), 705-730.

New B., Pallier C., Ferrand L., Matos R. (2001). Une base de données lexicales du français contemporain sur Internet: Lexique. L'année Psychologique, 101(3), 447-462.

ANNEXE : LA BASE DE DONNEES CHACQFAM

La base de données peut être obtenue à l'adresse http://psycholinguistique.unige.ch/

NOTE D'AUTEUR

Affiliation

Laboratoire de Psycholinguistique Expérimentale,

Faculté de Psychologie et des Sciences de l'Éducation,

Université de Genève, GE, Suisse.

Remerciements

Ce travail a pu être produit grâce à l'apport de Messieurs F. Serena et P. Demierre (Service Développement de la Division Informatique de l'Université de Genève), qui ont participé à la conception du système, et en ont assumé entièrement le développement. Je tiens à les remercier, ainsi que tous les anonymes qui ont volontairement participé à cette enquête et nous ont offert du temps et de l'information.

Je remercie également le Pr. U. H. Frauenfelder, qui m'a permis de réaliser ce travail dans son laboratoire, à Genève.

Contact

Christian.Lachaud@pse.unige.ch

Dr. Christian LACHAUD

Laboratoire de Psycholinguistique Expérimentale

FPSE, Université de Genève

Boulevard du Pont d'Arve, 40

CH 1205 Genève, GE - SUISSE

TABLEAUX

Tableau I

Descriptif des caractéristiques moyennes des mots de la base de données CHACQFAM – Mean characteristics of CHACQFAM's vocabulary

	N			PU		N voisins		Fréquence	
	syllabe	lettre	phonème	orthogr	phonol	orthogr	phonol	Frantext	Fastsearch
Moy.	1.20	4.76	3.34	4.43	3.18	7.16	16.64	291.89	22210.43
Min.	1	1	1	0	0	0	0	0.03	11.49
Max.	2	10	8	10	8	46	40	23889	734542.05
E.-T.	0.40	1.54	1.26	1.70	1.37	5.84	10.08	1634.91	70692.85

Note. N pour nombre d'items, PU pour point d'unicité, orthogr pour orthographique, phonol pour phonologique, Moy. pour moyenne, Min. pour minimum, Max. pour maximum, E.-T. pour écart-type.

Tableau II

Résultat des tests de significativité - Results of the significativity tests

	$\chi^2_{(1, N = 1225)}$	p <
AoA	1.484	.23
Cert	18.326	.0001
Fam	0.346	.56
AoA.Cert	2.241	.14
AoA.Fam	0.774	.38
Cert.Fam	2.556	.11
AoA.Cert.Fam	2.042	.16

Note. AoA pour âge d'acquisition estimé ; Cert pour degré de certitude pour l'estimation de l'âge d'acquisition ; Fam pour degré de familiarité.

Tableau III

Comparaison statistique des données fournies par CHACQFAM avec celles fournies par 4 bases de données de référence - Statistical comparison between CHACQFAM's data and those provided by the 4 reference data bases

		Alario et col.	Ferrand et col.	Bonin et col.	Chalard et col.
CHACQFAM	N mots	49	86	59	36
	N rép	19.82	20.88	20.20	20.14
	%	4	7.02	4.82	2.94
AoA	β	0.84	0.84	0.76	0.74
	R^2 ajusté	0.7	0.7	0.58	0.54
Fam	β	0.66	-	0.3	-
	R^2 ajusté	0.43	-	0.07	-

Note. N mots pour le nombre de mots de CHACQFAM en commun avec chaque base comparée, N rép pour la taille moyenne de l'échantillon par mot dans CHACQFAM, % le pourcentage de chevauchement entre CHACQFAM et chacune des 4 bases. β est le coefficient de régression, et R^2 le coefficient de détermination.

Tableau IV

Comparaison des scores d'âge d'acquisition fournis par les 4 bases de données de référence - Comparing the scores of age of acquisition between the 4 reference data bases

N / R^2 ajusté	Alario et col.	Ferrand et col.	Bonin et col.	Chalard et col.
Alario et col.		28	6	226
Ferrand et col.	0.94		21	21
Bonin et col.	--	0.91		2
Chalard et col.	0.29	0.11	--	

Note. N est le nombre de mots que les bases ont en commun 2 à 2, et R^2 est le coefficient de détermination.

LEGENDE DES FIGURES

Figure 1. Distribution du nombre de mots en fonction de l'âge d'acquisition estimé (courbe : distance pondérée des moindres carrés) - Distribution of the number of words according to the estimated age of acquisition (curve: outdistance balanced least squares)

FIGURES

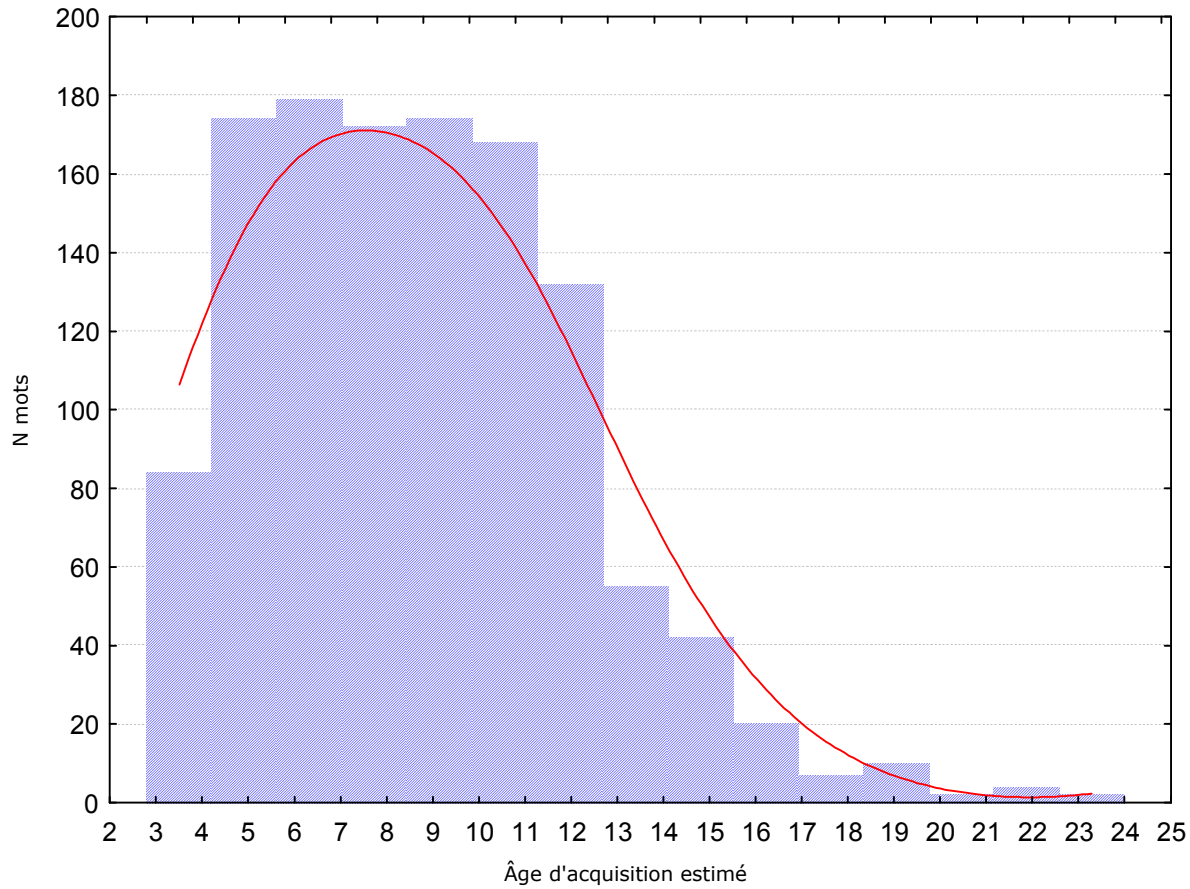

Figure 1.