Le traitement des mots singuliers et pluriels en français et en anglais

PAR BORIS NEW

Université Paris V
Institut Henri Pieron
71 avenue Edouard Vaillant
Tél: 01 55 20 58 64
E-mail : boris.new@univ-paris5.fr

Situation actuelle

Maître de conférence à l'université Paris V (Laboratoire de Psychologie Expérimentale - UMR 8581).

Diplômes et parcours universitaire

2004 : **Post-doctorat dans le cadre du projet Technolangue.** Laboratoire de Psychologie expérimentale. UMR CNRS 8581 - Université René Descartes, Paris V.

Beaucoup de mots utilisés quotidiennement sont des variantes d'autres mots, obtenus soit en combinant d'autres mots, soit en ajoutant un suffixe ou un préfixe à une racine existant déjà (chiens, chien**ne**, fille**tte**, brusque**ment**). Une question importante afin de savoir comment nous arrivons à lire est de savoir comment de tels mots sont

2003 : Boursier post doctoral de la Fondation FYSSEN. Sous la supervision de Marc Brysbaert Royal Holloway University of London, Royaume-Uni.

1998-02 : Thèse de Doctorat de Psychologie Cognitive. "La base Lexique et l'étude expérimentale de la flexion nominale en français" Thèse effectuée sous la direction de Juan Segui. Laboratoire de Psychologie expérimentale - UMR CNRS 8581 - Université René Descartes, Paris V.

1997-98 : DEA de Psychologie des Processus cognitifs. Université Paris 8 Vincennes - Saint-Denis.

Publications

New, B., Brysbaert, M., Segui, J., Ferrand, L., Rastle, K. (Sous Presse) The Processing of singular and plural nouns in French and English. *Journal of Memory and Language.*

New, B., Pallier, C., Brysbaert, M., Ferrand, L. (Sous Presse) Lexique 2.5 : A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers.*

Rastle, K., Davis, M., New, B. (Sous Presse) Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review.*

Alario, F-X., Ferrand, L., Laganaro, M., New B., Frauenfelder, U., & Segui, J. (Sous Presse) Predictors of Picture Naming Speed. *Behavior Research Methods, Instruments, & Computers.*

reconnus. Dans cet article, nous nous limiterons à la question de savoir comment les mots suffixés sont reconnus. Les mots suffixés peuvent être soit des mots fléchis soit des mots dérivés de (mots) racines. Les mots fléchis sont des variations de la forme du mot original conservant la catégorie grammaticale du mot d'origine et ne changeant pas ou peu le sens du mot. Pour les verbes, ses formes fléchies sont ses formes conjuguées. Pour les noms, ses formes fléchies sont ses formes plurielles et ses formes féminines. Les mots dérivés permettent la formation d'un mot nouveau et changent souvent la signification et la classe grammaticale du mot d'origine (**més**entente, re**vivre**).

Classiquement, les mots suffixés peuvent être compris de deux façons : soit ils sont stockés en mémoire et reconnus de façon globale (c'est la théorie du stockage exhaustif (Butterworth, 1983)), soit ils sont décomposés et reconnus par le biais de leurs différents morphèmes (re+vivre, més+entente ; c'est la théorie décompositionnelle (Taft, 1979 ; Taft et Forster, 1975)). Différentes théories attachent une importance différente à ces deux voies. Ainsi les modèles connexionnistes du traitement morphologique postulent qu'avec la présentation répétée de mots fléchis et dérivés, des réseaux associatifs vont se former entre les mots morphologiquement reliés. Dans ces modèles (p.ex., Rumelhart & McClelland, 1986), les relations morphologiques sont secondaires, puisqu'elles sont dérivées de l'association entre les formes de surface des mots et leur signification.

D'autres modèles mettent plus l'accent sur la décomposition et postulent que les mots complexes sont accédés prioritairement par leurs morphèmes. En effet, un certain nombre de mots suffixés peuvent être reconnus en appliquant une règle simple (mot + s pour les formes plurielles d'un mot, + ment pour former des adjectifs, etc.). Pour ces formes, il n'y aurait nullement besoin de les stocker séparément puisqu'elles peuvent être facilement comprises en appliquant les bonnes règles. Seules les exceptions à ces règles auraient besoin d'être stockées (yeux). Un tel modèle a d'ailleurs été proposé par Taft (p.ex. Taft, sous presse) pour le traitement des mots en anglais et par Clahsen (p.ex. Clahsen, 1999) pour le traitement des mots en allemand.

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

Ferrand, L., & New, B. (2003). Syllabic length effects in visual word recognition and naming. *Acta Psychologica,* 113, 167-183.

New, B., Pallier, C., Ferrand, L., Matos, R. (2001) Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Pschologique,* 101, 447-462.

Un troisième groupe de modèles, postule que les mots complexes morphologiquement réguliers peuvent être à la fois stockés et décomposés. Parmi ces modèles à deux routes, ceux-ci diffèrent grandement de par l'importance qu'ils accordent à une route plutôt qu'à une autre.

Par exemple, Pinker et Ullman (2002) défendent un modèle à deux routes pour le traitement des formes passées des verbes en anglais. Toutefois, bien qu'ils disent qu'un mot régulier peut être stocké individuellement dans le lexique mental, le reste de l'article nous fait bien comprendre que la route lexicale n'est pas très importante pour la reconnaissance et la production de ces formes. À l'inverse, Caramazza, Laudanna, et Romani (1988) considèrent que la route lexicale est généralement la plus rapide (même pour les formes régulières) et que la route décompositionnelle est seulement importante pour les mots nouveaux ou très rares.

Un autre modèle à deux voies qui se situe à l'intersection de ces deux extrêmes est celui de Baayen, Schreuder et leurs collègues (Baayen, Dijkstra, & Schreuder, 1997). Selon ce modèle, les routes de stockage et de décomposition sont activées en parallèle et leur contribution relative pour la reconnaissance des mots complexes dépend d'une série de facteurs. Pour les mots suffixés, Bertram et al. (2000) postulent l'influence de trois facteurs : le type de mot, la productivité du suffixe et si le même suffixe est utilisé dans plus d'un type de dérivation ou de flexion.

Baayen et al. (1997) ont ainsi proposé un modèle à routes parallèles constitué de trois niveaux. Lors du premier niveau, la chaîne de caractères active un grand nombre de représentations stockées en mémoire à long terme. En général, cela comprend le mot en entier mais aussi, en parallèle, les segments du mot formant des unités de sens. Ainsi un stimulus tel que *dogs* activera non seulement la représentation de *dogs,* mais aussi celle de *do, dog,* et *-s.*

Les représentations qui atteignent un certain seuil passent alors au deuxième niveau de traitement. À ce niveau, un processus de vérification a lieu pour les unités qui sont plus courtes que le mot lui-même. Ce processus de vérification s'assure que la combinaison d'unité morphémique est bien autorisée grammatica-

lement. Enfin, au dernier niveau, les traits sémantiques et syntaxiques des unités sont activés. Pour les différentes combinaisons d'unités, cela implique le calcul de sa signification.

Comme les mots activent simultanément les représentations qui correspondent au mot et celles correspondant aux morphèmes, le modèle de Baayen et Schreuder comprend à la fois une route lexicale et une route décompositionnelle. La vitesse de traitement de chacune des deux routes dépendra à la fois de la fréquence du mot et de la fréquence des morphèmes augmentée des temps de segmentation, vérification et composition.

Pour les pluriels, la fréquence de la route lexicale dépend de la fréquence de surface de la forme plurielle. Ainsi, la route lexicale sera plus rapide pour un mot de haute fréquence plurielle tel que *nuages* que pour un mot de basse fréquence plurielle tel que *piano.* Pour les singuliers, la vitesse de la route lexicale dépendra de la somme des fréquences des mots au singulier et au pluriel (fréquence cumulée) car les singuliers sont activés lors de la présentation de formes au singulier mais aussi lors de la présentation de formes plurielles. La vitesse de décomposition dépend donc de la fréquence des segments constitutifs mais aussi du coût de la décomposition. Ainsi pour les noms avec un pluriel de haute fréquence et un singulier de basse fréquence (p.ex. *nuages*), la route lexicale sera généralement plus rapide car celle-ci dépend de la fréquence de surface du pluriel et la route décompositionnelle de la fréquence cumulée et du temps de décomposition. Pour ce type d'item, les temps de reconnaissance dépendront donc beaucoup de la fréquence des formes plurielles. En revanche les mots ayant des pluriels de basse fréquence auront plus de chance d'être reconnus par la route de décomposition et, de ce fait, d'être plus sensibles à la fois à la fréquence cumulée et au temps de décomposition.

Des données semblant accréditer cette hypothèse ont été trouvées concernant le hollandais et le finlandais. Cependant des données contradictoires ont aussi été trouvées en anglais par Sereno et Jongman (1997). Dans une première expérience Sereno et Jongman ont présenté des mots de même fréquence cumulée mais étant soit singuliers dominants *(faiblesse)* soit pluriel dominant *(chaussure)*. Ils ont trouvé que pour les pluriels dominants, les formes plurielles sont traitées plus rapidement que les singuliers. À l'inverse, pour les singuliers dominants, ils observent que les formes au singulier sont traitées plus rapidement que les formes plurielles.

Enfin, ils présentent dans une dernière expérience des mots de même fréquence de surface mais ayant des fréquences cumulées haute et basse. Les performances des sujets ne varient pas selon les deux groupes d'items. Sereno et Jongman en concluent que les temps de décision lexicale pour les noms en anglais dépendent uniquement de leur fréquence de surface et donc que leurs résultats sont en faveur de l'existence d'un modèle à stockage exhaustif.

Ces résultats sont clairement en contradiction avec les données obtenues par Schreuder et Baayen en hollandais. Elles sont aussi en contradiction avec leur modèle. Pour cette raison, nous avons voulu voir quel modèle se généraliserait le mieux à une autre langue telle que le français.

Une des caractéristiques intéressantes du pluriel en français vient du fait que ce paradigme est très similaire au pluriel en anglais. En effet, la terminaison *-s* est une terminaison très régulière et productive (plus de 98 % des pluriels des adjectifs et des noms finissent en *-s*). Il dispose d'un rival (la deuxième personne du singulier) mais la fréquence de ce rival est bien moindre. Selon Bertram et al. (2000), ces caractéristiques impliquent que les pluriels devraient être majoritairement retrouvés grâce à la voie décompositionnelle plutôt que par la route lexicale.

EXPÉRIENCE 1
(manipulation de la fréquence de surface)

Étant donné les résultats conflictuels obtenus en hollandais par Baayen, Dijkstra et Schreuder, nous avons voulu tester dans notre première expérience l'influence de la fréquence de surface en français. Plus spécifiquement, nous avons voulu étudier l'importance de la fréquence de surface dans le traitement des formes au singulier ou au pluriel. Pour cela, nous avons créé deux listes de mots ayant la même fréquence cumulée mais

différant de par leur fréquence de surface. La moitié des stimuli était singulier dominant (mot ayant une fréquence du singulier supérieure à leur fréquence du pluriel) alors que l'autre moitié était composée de mots pluriel dominant. Les mots étaient présentés dans une tâche de décision lexicale visuelle. Cette tâche consiste pour un sujet à déterminer le plus vite possible lors de l'apparition d'une chaîne de caractères à l'écran si cette chaîne constitue un mot ou un non-mot.

Matériel

Les stimuli sont 48 noms extraits de la base de données Lexique (New, Pallier, Ferand, & Matos ; 2001) qui est une nouvelle base de données pour les mots dont la fréquence est basée sur large corpus de textes récents (31 millions de mots). Avant la parution de cette base de données, les études concernant des formes fléchies étaient très difficiles car les bases de données existantes ne donnaient pas la fréquence des formes fléchies (p.ex. *chienne* ou *dansait*). De plus, tous les noms prennent leur pluriel en *-s.*

La première liste de 24 mots était composée de mots ayant un singulier plus fréquent que leur pluriel (mots singulier dominant). La deuxième liste était composée de 24 mots ayant un pluriel plus fréquent que leur singulier. La fréquence cumulée ne différait pas entre les deux listes. Les mots des deux listes étaient aussi appareillés sur leur nombre de lettres et leur nombre de syllabes.

48 non-mots ont aussi été créés en remplaçant une voyelle par une voyelle ou une consonne par une consonne à des mots de la langue française. Ces non-mots étaient prononçables et respectaient les contraintes phonotactiques de la langue française. La moitié des non-mots finissaient en *-s.*

Procédure

Chaque stimulus (mot ou non-mot) était présenté à l'écran et le sujet devait répondre le plus vite possible en faisant le moins d'erreur possible.

Résultats

Les résultats sont présentés ci-dessous en haut à droite dans la figure 1.

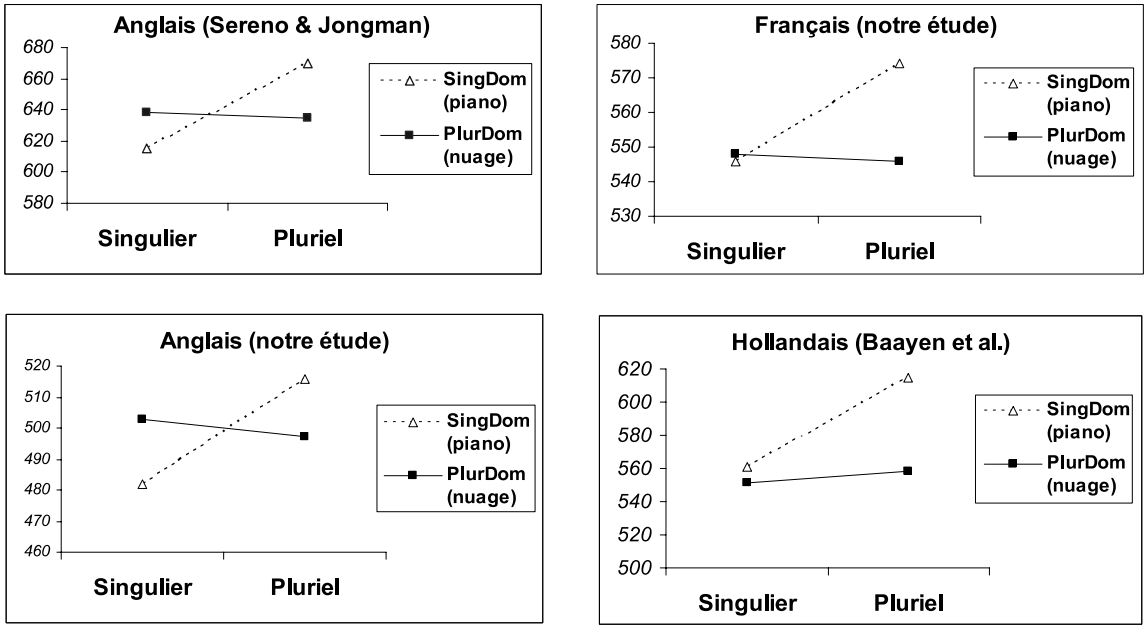

Figure 1 : Temps de décision lexicale des mots singulier dominant ou pluriel dominant présentés au singulier ou au pluriel en anglais, en français et en hollandais.

La question que nous nous posions lors de la réalisation de cette expérience était de savoir à quel point les temps de décision lexicale aux noms pluriels et singuliers étaient déterminés par leur fréquence de surface quand leur fréquence cumulée était contrôlée.

Nous constatons que nos résultats sont extrêmement similaires à ceux de Baayen et al. en hollandais. Pour les items singuliers dominants, les items au singulier sont traités significativement plus rapidement que les pluriels alors qu'aucune différence n'est observée pour les items pluriels dominants.

Pour les items singuliers dominants, cela signifie que la fréquence cumulée n'est pas le seul déterminant des temps de réaction mais que -soit la fréquence de surface joue et les pluriels sont plus lents car ils sont moins fréquents, -soit il y application de règle et les pluriels mettent plus de temps à être traités en raison de ce traitement supplémentaire. Pour les items pluriels dominants, cela signifie que la fréquence de surface n'est pas déterminante lors de leur traitement. Ces deux résultats permettent d'ores et déjà d'écarter l'hypothèse des modèles exclusivement à stockage exhaustif ou exclusivement à décomposition.

Après avoir exploré dans cette expérience si les mots en français étaient sensibles à leur fréquence de surface, nous avons voulu savoir dans une deuxième expérience si les temps de réaction pour les items au singulier étaient influencés par leur fréquence cumulée.

EXPÉRIENCE 2

Afin d'étudier l'influence de la fréquence cumulée sur les items singuliers, nous avons comparé les temps de réaction à des mots ayant la même fréquence de surface mais ayant des fréquences de la forme plurielle soit très basses *(neige)*, soit très hautes *(doigt)*.

Matériel et procédure

44 mots ont été sélectionnés à partir de *Lexique* et 44 non-mots ont été créés suivant la même procédure exposée pour l'expérience 1. Une première liste de 22 mots a été créée avec une fréquence du singulier de 16 et une fréquence du pluriel de 4.

Résultats et discussion

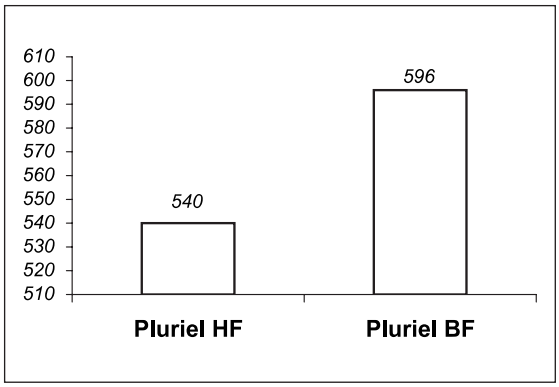

Figure 2 : Temps de décision lexicale en français pour des items de fréquence de surface identique mais dont la fréquence des pluriels diffère.

Les résultats montrent que quand deux items au singulier ont des fréquences de surface comparables mais des fréquences cumulées différentes, la forme au singulier ayant le pluriel le plus fréquent est traitée plus rapidement. Ces résultats sont complètement cohérents avec les résultats obtenus en hollandais mais diffèrent de ceux obtenus en anglais par Sereno et Jongman.

EXPÉRIENCE 3

En regardant attentivement les données de Sereno et Jongman en anglais, nous nous sommes aperçus de certaines différences méthodologiques entre leurs expériences et celles réalisées en français et en hollandais. Leurs expériences ont tout d'abord utilisé des listes bloquées d'items soit uniquement au singulier, soit uniquement au pluriel. Cette présentation bloquée aurait pu encourager les sujets à ignorer la terminaison -s durant l'expérience. Un autre problème provient du fait que Sereno et Jongman ont utilisé des fréquences inspirées du corpus de Brown basées sur 1 million de mots, ce qui est peu comparé aux corpus utilisés en français (31 millions de mots) et en hollandais (42 millions de mots). Pour cette raison, nous avons décidé de répliquer l'expérience de Sereno et Jongman en utilisant la même procédure que nous avons utilisé en français.

Matériel

Nous avons sélectionné deux listes de 24 noms à partir de la base de données Celex (Baayen, Piepenbrock, & van Rijn 1993) basé sur un corpus de 17,7 millions de mots. La première liste était composée d'items singuliers dominants avec une fréquence moyenne de 25 par million pour les singuliers et de 8 par million pour les pluriels. La deuxième liste était composée de pluriels dominants avec des fréquences respectives de 9 et 26 occurrences par million respectivement. La fréquence de base, le nombre de lettres et le nombre de syllabes ne différaient pas entre les deux listes.

En outre, nous avons créé 48 stimuli non-mots phonotactiquement légaux et ayant en moyenne le même nombre de lettres, de syllabes et de voisins. La moitié des non-mots finissaient en *-s*.

Procédure

La procédure était la même que dans l'expérience 1.

Résultats

Comme vous pouvez le constater dans la figure 1 en bas à gauche, le pattern trouvé en anglais est assez intrigant : d'une certaine façon il est très similaire au pattern obtenu en français ; en effet, il y a une différence significative des temps de réponse entre les items singuliers et les items pluriels pour les singuliers dominants mais pas pour les pluriels dominants. D'autre part, nous obtenons un pattern proche de celui obtenu par Sereno et Jongman. Si nous regardons les temps pour les formes au singulier, les temps de réaction sont plus rapides pour les singuliers dominants que pour les pluriels dominants. Et si nous regardons les temps pour les formes au pluriel, les temps de réaction sont plus rapides pour les pluriels dominants que pour les singuliers dominants.

EXPÉRIENCE 4

Sereno et Jongman n'ont pas trouvé d'effet de la fréquence cumulée sur les noms au singulier en anglais. Une explication à leur résultat pourrait être qu'ils ont utilisé des mots au singulier de très haute fréquence (95 occurrences par million) alors que les études en hollandais et en français ont utilisé des mots de fréquence moyenne (10 à 15 occurrences par million). Il est donc tout à fait possible que Sereno et Jongman n'ont pas trouvé d'effet de la fréquence cumulée car leurs items au singulier avaient déjà un effet plafond et ne pouvaient pas profiter d'activation additionnelle de la part des pluriels. Pour cela nous avons donc sélectionné deux listes de mots au singulier ayant une fréquence de surface moyenne.

Matériel

48 mots ont été sélectionnés à partir de Celex. La première liste de 24 mots avait un pluriel de haute fréquence (fréquences de 15 et 39 pour les singuliers et les pluriels respectivement). La deuxième liste était composée de 24 noms de basse fréquence (fréquences de 16 et 1,9). Les deux listes de mots étaient contrôlées sur les mêmes paramètres que dans l'expérience 1.

Résultats

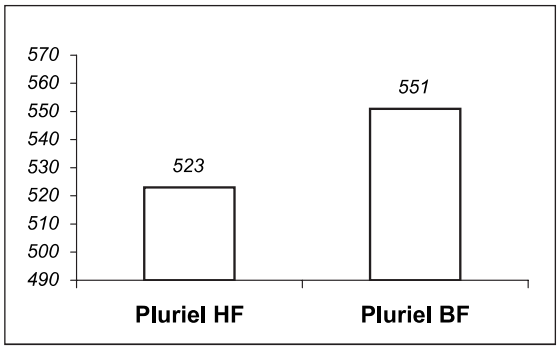

Figure 3 : Temps de décision lexicale en anglais pour des items de fréquence de surface identique mais dont la fréquence des pluriels diffère.

Comme vous pouvez le constater sur la figure 3, dans cette expérience, nous montrons qu'en anglais, les temps de décision lexicale pour les mots au singulier sont influencés par la fréquence de leur pluriel tout comme en hollandais et en français.

Discussion

Dans cette recherche, nous avons cherché à déterminer comment les mots présentés visuellement au singulier ou au pluriel sont reconnus. Si l'on se fie à la figure 1, nous constatons trois phénomènes :

1) En anglais, en hollandais et en français, les temps de réaction aux mots singuliers sont influencés par la fréquence de leur forme au pluriel. (Expérience 2 et 4).

2) Dans ces trois langues, pour les items singuliers dominants, les temps de réaction aux formes plurielles sont plus lents que les temps de réaction aux formes au singulier.

3) Dans ces trois langues, pour les items pluriels dominants, les temps de réaction aux formes au singulier sont identiques à ceux des formes plurielles.

Le premier et le troisième point nous permettent d'éliminer définitivement l'hypothèse de modèles exclusifs qui utilisent uniquement un stockage exhaustif ou des règles de décomposition. En effet selon un modèle à stockage exhaustif, pour les items pluriels dominants, les formes plurielles devraient être traitées plus rapidement que les formes au singulier, ce qui n'est pas le cas dans aucune des langues.

En revanche, selon un modèle à décomposition obligatoire, les items présentés au pluriel devraient systématiquement donner lieu à des temps de réaction plus longs. En effet, dans ce cas les formes plurielles nécessiteraient systématiquement l'emploi d'une procédure que l'on peut supposer coûteuse en temps de décomposition.

Ces résultats obtenus dans diverses langues valident donc, en ce qui concerne le traitement des formes plurielles régulières, les modèles à deux voies alliant route décompositionelle et route lexicale. Plus spécifiquement nos résultats peuvent être expliqués par le modèle de Schreuder et Baayen qui fait le postulat original que la présentation d'un item pluriel activerait son correspondant singulier alors que l'inverse ne serait pas vrai ce qui rendrait, à terme, les items singuliers sensibles à leur fréquence cumulée tandis que les items pluriels pourraient, suivant leurs propriétés, être davantage sensibles à leur fréquence de surface ou à leur fréquence cumulée.

BIBLIOGRAPHIE

BAAYEN, R. H., DIJKSTRA, T., & SCHREUDER, R. (1997). Singulars and plurals in Dutch : Evidence for a parallel dual-route model. *Journal of Memory and Language,* 37, 94-117.

BERTRAM, R., SCHREUDER, R., & BAAYEN, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 489-511.

BUTTERWORTH, B. (1983). Lexical representation. In B. Butterworth (Ed.), Language *Production Volume 2 : Development, Writing and Other Language Processes.* (pp. 257-294). London : Academic Press.

CARAMAZZA, A., LAUDANNA, A., & ROMANI, C. (1988). Lexical access and inflectional morphology. *Cognition,* 28, 297-332.

CLAHSEN, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences, 22,* 991-1060.

PINKER, S., & ULLMAN, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Science, 6,* 456-463.

TAFT, M. (in press). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology.*