Article

« Connexionisme et attribution du genre en français : modèle d'acquisition ou de classification? »

Patricia Rodrigues et Robert Boivin

Revue québécoise de linguistique, vol. 28, n° 2, 2000, p. 29-50.

Revue québécoise de linguistique, vol. 28, nº 2, 2000, © RQL (UQAM), Montréal

CONNEXIONISME ET ATTRIBUTION DU GENRE EN FRANÇAIS : MODÈLE D'ACQUISITION OU DE CLASSIFICATION ?*

Patricia Rodrigues et Robert Boivin
Université du Québec à Montréal

1. Introduction

Depuis le début des années quatre-vingt, avec le développement des réseaux de neurones artificiels, un nouveau paradigme appelé *connexionisme* vient s'opposer aux approches symboliques et représentationnelles du cognitivisme classique. Le connexionisme se veut une solution aux difficultés théoriques et pratiques rencontrées par le cognitivisme classique, notamment sur le plan de l'implémentation d'un modèle capable de représenter les fonctions cognitives. Il se distingue du paradigme symbolique en proposant une conception de la cognition qui ne met pas en jeu la manipulation de symboles au moyen de règles. En s'inspirant de la structure en réseaux de neurones du cerveau, il propose un modèle computationnel nouveau et très puissant de type non symbolique représenté par un réseau d'unités élémentaires connectées entre elles. On dit de ce réseau, qui opère en parallèle et de façon interactive, qu'il est capable d'apprendre avec l'expérience. Les connaissances ainsi acquises sont emmagasinées de façon distribuée dans la totalité du réseau[1].

Au cœur du débat entre le connexionisme et le cognitivisme classique, nous trouvons les questions portant sur l'analyse du fonctionnement du langage. Plusieurs modèles connexionistes ont été développés pour modéliser le traitement du langage. Le connexionisme fournit un modèle qui permet d'explorer l'hypothèse de l'émergence des représentations linguistiques de l'activité neuronale; ces représentations apparaissent alors comme une conséquence de l'organisation

* Les auteurs remercient le professeur Tom Cobb, du Département de linguistique et de didactique des langues de l'Université du Québec à Montréal.

1 Pour plus de détails sur le cadre connexioniste, nous suggérons au lecteur les ouvrages suivants : Fausett 1994, Haykin 1999, Mehrotra, Mohan et Ranka 2000.

neuronale. Pour les connexionistes, bien que le comportement langagier puisse être décrit au moyen de règles, cela ne veut pas dire qu'il est régi par ces règles. Les réseaux connexionistes peuvent simuler un comportement basé sur des règles sans pour autant avoir ces règles explicitées dans leur système, comme c'est le cas pour les systèmes symboliques.

Plusieurs études ont utilisé les réseaux connexionistes pour modéliser les processus d'acquisition ou de traitement du langage. Nous pouvons citer, par exemple, celles qui concernent la formation du passé en anglais (Rumelhart et McClelland 1986, Plunkett et Marchman 1991). Elman et coll. 1996 présentent aussi plusieurs modèles créés pour rendre compte du traitement du langage, tels que le «single recurrent network» qui prédit les continuations possibles dans une séquence donnée de mots; pour les auteurs, le modèle est capable d'apprendre les catégories de noms et de verbes sur la base de la représentation distributionnelle. Ces modèles ont toutefois soulevé beaucoup de controverses et de critiques. Par exemple, Marcus 1998 critique les systèmes modélisant l'acquisition du passé parce qu'ils ne peuvent généraliser leurs connaissances à des nouveaux mots qui ne ressemblent pas phonologiquement à ceux avec lesquels le réseau a été entraîné[2].

Parmi ces études sur le connexionisme et le traitement du langage, nous trouvons celle de Sokolik et Smith 1992, qui décrit un système connexioniste développé pour rendre compte de l'assignation du genre en français. Des études sur l'acquisition du genre par les locuteurs natifs du français montrent qu'il existe, outre l'indication contextuelle (genre du déterminant, genre des adjectifs), deux différents types d'indications pouvant amener une classification par le genre. Le genre peut être déduit de la forme phonétique des terminaisons (p. ex. [o] est une terminaison masculine plus de neuf fois sur dix), puis par la composition morphologique des noms (p. ex. adjectif+*(i)té* donne un mot féminin) (Tucker, Lambert et Rigault 1977). Sokolik et Smith prétendent que ces indices phonologiques et morphologiques sont suffisants pour l'assignation correcte du genre par les locuteurs du français; de plus, ils suggèrent que ce processus ignore la formulation de règles d'inférence et se produit au niveau perceptuel.

Dans notre recherche, nous soumettons un corpus de français oral d'enfants de 5 à 8 ans encodé phonétiquement à un réseau neuronal à trois couches afin de vérifier si le modèle connexioniste peut véritablement rendre compte de l'assignation du genre dans les cas susmentionnés sans utilisation de règles. Dans les prochaines sections, nous décrirons d'abord l'étude de Sokolik et Smith, pour ensuite présenter les études sur le genre en français qui encadrent

2 Le concept d'*entraînement* est utilisé en connexionisme pour faire référence à la capacité d'apprentissage du système.

notre recherche. À la fin, nous décrirons la simulation mise en place pour atteindre nos objectifs et discuterons de ses résultats.

2. L'étude de Sokolik et Smith 1992

Pour Sokolik et Smith 1992, la compétence des apprenants de français langue seconde (ainsi que celle des locuteurs natifs) dans l'identification du genre grammatical des noms reflète leur habileté à reconnaître les patrons phonologiques et morphologiques constituant des indices de genre. Les auteurs attribuent toutefois la détection de ces patrons à des processus de perception de niveau inférieur («low-level perceptual inputs») et non à des inférences réalisées en tenant compte des informations disponibles à partir du contexte ou présentes dans la mémoire du locuteur. Autrement dit, pour eux, l'information inhérente à la structure des noms français est suffisante pour permettre l'assignation correcte du genre sans l'utilisation d'autres types d'information.

Pour démontrer que le genre en français peut être appris sur la base d'un traitement perceptif de niveau inférieur, ils ont réalisé une simulation en utilisant un modèle de type connexioniste connu comme «pattern associator». Ce type de réseau est utilisé pour grouper des classes d'information d'entrée en des codes communs à chaque classe. Une couche d'unités d'entrée représentant l'information à classer est connectée directement à une couche d'unités de sortie représentant les classes. Le réseau est capable d'extraire des traits abstraits communs à l'ensemble des entrées et, sur la base de ces traits, d'associer l'entrée à l'une des classes représentées dans la sortie.

Sokolik et Smith ont utilisé comme corpus une liste de 600 noms (de trois à huit lettres) dont 450, sélectionnés arbitrairement, ont servi à l'entraînement du réseau. Les 150 noms restants ont été utilisés pour tester la capacité du modèle à généraliser ses connaissances, c'est-à-dire à identifier le genre des noms qu'il ne «connaissait» pas.

Les mots soumis au réseau ont été encodés à partir de leur forme écrite à l'aide du paradigme «lettre / position dans l'alphabet», où chaque lettre du mot nécessite 28 bits pour être encodée (les 26 lettres de l'alphabet plus deux bits pour le «è» et le «é»). Pour chacune des lettres, le bit correspondant à sa position dans l'alphabet est activé. C'est le type d'encodage des mots qui a défini le nombre d'unités présentes dans la couche d'entrée : le mot le plus long ayant huit lettres, la couche d'entrée est donc constituée d'un vecteur de 224 unités. Les mots ayant moins de huit lettres sont alignés à droite. La règle d'apprentissage appliquée au modèle est connue comme règle delta, une règle qui utilise la

différence entre l'activation désirée et l'activation obtenue dans l'unité de sortie pour changer les poids des connexions et guider l'apprentissage du réseau.

Après cinq cycles d'entraînement, le réseau a été capable de classer correctement 86,2 % des noms. Ensuite, le réseau a été testé avec l'ensemble des 150 mots réservés pour vérifier sa capacité à généraliser ses connaissances. L'analyse des résultats montre que le modèle a été capable de prédire correctement le genre de 76 % des nouveaux mots.

Les auteurs concluent qu'un modèle connexioniste simple est capable d'apprendre le genre d'une bonne partie des noms français sans se baser sur les informations issues du contexte (l'accord avec l'article ou l'adjectif) ni sur la signification du nom, et sans avoir été programmé avec des règles spécifiques concernant la formation du genre. De plus, pour les auteurs, ce processus d'apprentissage est réalisé sur la base d'un traitement perceptif de niveau inférieur.

3. Critique de Sokolik et Smith 1992

Dans une critique de l'étude de Sokolik et Smith, Carrol 1995 affirme que ces conclusions ne découlent pas des résultats obtenus dans la simulation. Elle critique d'abord ce que les auteurs appellent le traitement perceptif de niveau inférieur. Selon Carrol, l'information lettre-plus-position ne correspond pas à un traitement de niveau visuel, mais à une construction plus abstraite. De plus, elle soutient que les auteurs ne peuvent pas conclure que le système a «appris» le genre en français à partir des résultats obtenus. Pour Carrol, le processus concernant l'apprentissage du genre est très complexe et doit tenir compte de plusieurs facteurs : d'abord, on doit apprendre que le français possède un système de genres (le modèle de Sokolik et Smith est forcé de classer l'input selon le genre); ensuite, on doit apprendre que le français a deux genres (le nombre de genres est stipulé a priori dans le modèle de Sokolik et Smith); on doit aussi apprendre que les noms possèdent un genre, mais que d'autres catégories n'en possèdent pas (l'input dans le modèle de Sokolik et Smith inclut seulement des noms). Elle conclut que, étant donné la quantité d'information linguistique cachée dans le design de l'expérimentation, on peut contester la conclusion que le système a «appris» l'association entre les noms et le genre.

Nous croyons pour notre part que la conclusion de Sokolik et Smith voulant qu'un réseau connexioniste puisse apprendre le genre en français à partir de la seule structure des mots semble précipitée. Dans leur travail, ils ne s'occupent pas de savoir sur quelle base le réseau a classé les noms. De plus, le choix de soumettre au système un corpus où les mots sont encodés à partir de leur forme

orthographique a pu largement fausser les données. En effet, plusieurs indices de genre peuvent se trouver dans l'orthographe sans toutefois être présents à l'oral (p. ex. la terminaison féminine en *-ée*). Ce choix méthodologique, ainsi que l'analyse incomplète des données, font en sorte que les résultats que les auteurs présentent peuvent être mis en doute.

4. But de la recherche

Le but du présent travail est de vérifier les conclusions de Sokolik et Smith sur l'apprentissage du genre en français au moyen d'un réseau connexioniste. Pour ce faire, nous avons réalisé une expérimentation utilisant un réseau connexioniste multicouches auquel nous avons présenté un corpus représentatif du langage parlé transcrit en symboles phonétiques. Une analyse approfondie des résultats obtenus nous permettra de connaître comment le réseau organise les données lorsqu'il effectue leur classement.

Contrairement aux suggestions de Carrol 1995, notre système sera préspécifié pour les genres grammaticaux du français tout comme le système de Sokolik et Smith 1992. Ce choix méthodologique nous permettra de focaliser sur la classification des mots plutôt que sur la capacité d'un système à créer des classes, ce qui dépasserait les buts que nous nous fixons dans le cadre de cette recherche. Nous demeurons toutefois d'accord avec Carrol.

5. Le genre grammatical en français

Le français possède un système de classification nominale appelé genre grammatical constitué de deux classes : le masculin et le féminin. Tous les noms communs en français doivent, invariablement, appartenir à une de ces deux classes. Le locuteur natif du français a peu de difficulté à maîtriser le genre des noms, ce qui n'est toutefois pas le cas des apprenants de français langue seconde, pour qui l'attribution du genre est souvent un casse-tête sans fin. Cette problématique a amené quelques chercheurs à se pencher sur l'étude de l'attribution du genre grammatical en français.

Selon Desrochers et coll. 1989, il y a deux classes d'indices qui peuvent potentiellement influencer l'identification du genre, soit les caractéristiques sémantiques du signifié et les caractéristiques structurelles du signifiant. Pour une bonne partie des noms caractérisés par le trait [+animé], les caractéristiques sémantiques du signifié sont significatives; dans ces cas, on observe une

correspondance claire entre le sexe du référent et le genre du nom (p. ex. *actrice, père*), bien que cette correspondance ne soit pas toujours parfaite (p. ex. *les étudiants, la sentinelle, la baleine*). Toutefois, dans le cas des noms caractérisés par le trait [-animé], la motivation sémantique est absente ou, tout au moins, très opaque.

En ce qui concerne la structure des mots, Tucker, Lambert et Rigault 1977 ont démontré, à partir d'une analyse statistique des noms du *Petit Larousse* (édition 1959), qu'il existe une relation entre la terminaison phonétique des noms et l'attribution du genre grammatical; par exemple, la probabilité que *–illon* termine un mot masculin s'élève à 95 %. Cependant, la valeur de prédiction des terminaisons phonétiques est très variable, plusieurs étant ambiguës, comme *–re* ou *–que,* respectivement dans 53 % et 56 % des mots masculins. Tucker, Lambert et Rigault ont aussi démontré que lorsque les locuteurs natifs attribuent un genre à un nom inconnu, ils le font en accord avec la proportion de noms masculins et féminins du lexique ayant la même terminaison que le nom en question. Par exemple, lors d'un test de classement, des noms inventés se terminant par *–oire* (51 mots masculins et 44 noms féminins dans le lexique), ont été classés masculins par environ la moitié des sujets participant au test.

Pour Surridge 1986, le fait que la terminaison phonétique ne soit pas un indice fiable dans la totalité des cas mène à la question de savoir comment « les francophones décident du genre grammatical des noms qui, pour chaque terminaison, constituent la minorité ». Pour l'auteure, « les locuteurs savent d'une part qu'il existe des règles phonétiques, mais ils savent aussi de toute évidence que d'autres règles s'appliquent ». Selon elle, il ne faut donc pas se baser uniquement sur la terminaison phonétique des noms pour expliquer l'attribution du genre grammatical; on doit aussi avoir recours à leur structure morphosyntaxique. Surridge 1985, 1986 a démontré que le genre grammatical des noms complexes en français (noms suffixés et composés) est attribué selon des règles simples basées sur la structure morphologique du mot. Selon elle, la structure morphologique des noms complexes permet d'obtenir une indication incontestable du genre grammatical, même si la terminaison phonétique est ambiguë. Par exemple, tous les noms composés basés sur un verbe, comme *portefeuille*, sont masculins, alors que tous les noms formés par un adjectif auxquels est ajouté le suffixe *–té* ou *–ité* sont féminins.

Selon Surridge 1993, la relation entre les différentes règles d'attribution du genre en français (phonétiques, morphologiques) est hiérarchique, et l'acquisition de ces règles par les locuteurs natifs obéit à un ordre chronologique. Cela veut dire que l'enfant intègre premièrement l'information phonétique, puis l'information morphologique, mais l'utilisation ultérieure de cette information lors

de l'exposition à un nouveau mot sera faite en considérant premièrement la morphologie et ensuite seulement la forme phonétique. Par ailleurs, Barbaud, Ducharme et Valois 1982 démontrent qu'en français parlé du Canada, entre autres variétés, l'initiale vocalique d'un mot peut être un facteur décisif pour l'assignation du genre.

6. Méthodologie

6.1 Approche méthodologique

L'expérience vise à mettre au point un réseau connexioniste capable de procéder à la classification d'une liste de mots du français selon le genre. Les mots, tirés d'un corpus représentatif, sont présentés tels quels, sans aucun indice de genre supplémentaire (p. ex. l'appartenance à une catégorie lexicale).

6.2 Le réseau connexioniste

6.2.1 *Tlearn*

Le réseau connexioniste que nous avons utilisé a été développé à partir du programme *Tlearn*. *Tlearn* est un simulateur de réseaux neuronaux qui permet la réalisation de divers types d'architectures d'apprentissage supervisé en utilisant une règle delta généralisée (règle de rétropropagation).

6.2.2 L'architecture du système

Le réseau connexioniste que nous avons dessiné comprenait 80 unités d'entrée, correspondant au nombre d'unités des vecteurs que nous avons utilisés lors de la codification du corpus que nous décrivons dans la section 6.3.

Contrairement à Sokolik et Smith 1992, qui, comme nous l'avons vu à la section 2, ont utilisé un système à deux niveaux, l'architecture de notre système incluait également 20 unités cachées. Les unités cachées ont comme avantage de pouvoir apporter une solution aux problèmes qui ne sont pas linéairement séparables. En d'autres mots, si une solution à un problème s'avère possible, les unités cachées en permettront l'élaboration. Les unités cachées agissent comme des extracteurs de traits. Il est donc permis de croire que si la solution à l'assignation du genre se trouve à l'intérieur des mots, par exemple dans les

terminaisons, les unités cachées pourront en extraire les traits et trouver ainsi une solution à l'assignation du genre.

Le nombre d'unités cachées choisi est quelque peu arbitraire. L'expérimentation est le seul moyen de savoir quel nombre d'unités permettra un rendement optimal du système : «There is no generally acknowledged criterion for selecting appropriate numbers of hidden units for an arbitrary problem. The modeller must, therefore, experiment with network capacities in order to find a configuration suited for the problem.» (Plunkett et Marchman 1991)

Afin de fixer le nombre d'unité cachées à vingt, nous avons effectué plusieurs préexpérimentations avec le corpus d'entraînement. Nous avons fait varier le nombre d'unités par tranches de dix entre dix et soixante. Nous en avons fixé le nombre en accord avec le pourcentage d'erreurs le plus bas.

Pour terminer, notre système était composé de deux unités de sortie, l'une pour le féminin, l'autre pour le masculin.

6.3 Corpus

Le corpus utilisé comprenait 1372 mots tirés d'un corpus de Préfontaine et Préfontaine 1968. Ce corpus représente un échantillonnage du vocabulaire oral fondamental des enfants de 5 à 8 ans recueilli au cours de l'année scolaire 1966-1967. Il est original dans le sens qu'il s'agit d'un corpus entièrement basé sur la langue orale, contrairement aux corpus similaires recueillis auprès d'adultes.

Le corpus de Préfontaine et Préfontaine est divisé en plusieurs listes. Nous avons utilisé une seule de ces listes (liste E). Elle inclut tous les mots recueillis auprès des enfants de 5 à 8 ans et est présentée par ordre de fréquence décroissante. Nous avons effectué quelques modifications à cette liste. Comme nous l'avons vu à la section 5, Desrochers et coll. 1989 soulignent que les noms caractérisés par le trait [+animé] sont identifiés pour le genre par leur valeur sémantique plutôt que par leur terminaison. Il nous a donc paru logique d'éliminer tout nom portant le trait [+animé], puisque l'encodage que nous avons choisi ne tient pas compte des traits sémantiques.

Nous retrouvons trois sortes de mots dans le corpus : mots simples, mots suffixés et mots composés. La proportion de mots suffixés est inférieure à la proportion de mots simples, mais supérieure à la proportion de mots composés. Le tableau 1 présente la liste des terminaisons dont la fréquence est égale ou supérieure à 5 dans le corpus[3].

3 Nous avons considéré les terminaisons [ɔ̃] et [sjɔ̃] comme étant distinctes l'une de l'autre. Leurs fréquences respectives de genre semblent justifier ce choix.

Tableau 1
Liste de fréquences des terminaisons du corpus

API	Masc	Fém	Total
-ɔ̃	85	11	96
-o	86	3	89
-e	49	26	75
-i	38	19	57
-ɛt	3	50	53
-ɛ̃	44	2	46
-t	4	38	42
-r	20	16	36
-ã	35	1	36
-ɛ	32	4	36
-œr	11	17	28
-yr	2	25	27
-sjɔ̃	0	25	25
-in	0	22	22
-ɛl	7	15	22
-wɑr	9	12	21
-l	12	9	21
-ɛr	13	7	20
-as	2	17	19
-d	5	13	18

API	Masc	Fém	Total
-al	12	6	18
-ãʒ	11	7	18
-ɑ	17	0	17
-jɛr	0	16	16
-ɛn	1	15	16
-u	15	0	15
-ɔl	4	10	14
-ãs	1	12	13
-ɛs	0	13	13
-ij	0	12	12
-a	9	3	12
-k	7	4	11
-ɔr	10	1	11
-œj	10	1	11
-ø	10	1	11
-wa	1	9	10
-s	2	8	10
-ir	7	3	10
-ar	7	3	10
-ɛj	5	5	10

API	Masc	Fém	Total
-iz	0	9	9
-il	4	5	9
-ad	0	9	9
-ɔt	0	9	9
-yl	1	7	8
-ur	8	0	8
-up	1	6	7
-ik	1	6	7
-aj	5	2	7
-əri	0	7	7
-uj	0	6	6
-m	2	4	6
-is	2	4	6
-ym	4	1	5
-øz	0	5	5
-ɛz	0	5	5
-ɑr	5	0	5
-ɛm	4	1	5
-ʃ	0	5	5

6.3.1 Encodage du corpus

Le corpus a premièrement été transcrit en symboles phonétiques. Les symboles choisis ne correspondent pas aux symboles phonétiques habituellement reconnus parce que *Tlearn* ne peut lire que les caractères ASCII[4]. Un vecteur d'unités binaires correspondant aux traits phonologiques de chaque phonème a ensuite été assigné à chacun des symboles. Nous avons tenté de restreindre au maximum le nombre de traits afin d'éviter toute redondance et afin que les vecteurs soient les plus courts possible. Tous les phonèmes sont décrits par les traits suivants : *voyelle*, *voisé*, *sonant* et *nasal*. Ils possèdent également deux traits de mode (*haut* et *bas* pour les voyelles, *constrictif* et *occlusif* pour les

4 «American Standard Code for Information Interchange» : caractères typologiques (alphabet latin) ayant une valeur adaptée au traitement informatique.

consonnes) et deux traits de lieu (*avant* et *arrière* pour les voyelles, *labial* et *coronal* pour les consonnes). Les tableaux 2a et 2b présentent la codification des phonèmes du corpus[5].

Tableau 2a
Codification des phonèmes du corpus (voyelles)

API	ASCII	TRAITS							
					mode		lieu		
		V/C	voisé	sonant	haut	bas	avant	arrière	nasal
i	i	1	1	1	1	0	1	0	0
e	&	1	1	1	1	1	1	0	0
ɛ	$	1	1	1	0	0	1	0	0
a	a	1	1	1	0	1	1	0	0
ɑ	@	1	1	1	0	1	0	1	0
ɔ	^	1	1	1	0	0	0	1	0
o	o	1	1	1	1	1	0	1	0
u	u	1	1	1	1	0	0	1	0
y	y	1	1	1	1	0	0	0	0
ø	Q	1	1	1	1	1	0	0	0
œ	e	1	1	1	0	0	0	0	0
ɑ̃	A	1	1	1	0	0	0	1	1
ɛ̃	E	1	1	1	0	0	1	0	1
ɔ̃	O	1	1	1	0	0	0	1	1
œ̃	U	1	1	1	0	0	0	0	1
j	j	0	1	1	1	0	1	0	0
ɥ	h	0	1	1	1	0	0	0	0
w	w	0	1	1	1	0	0	1	0

5 Certains choix de convention ont dû être faits afin de créer les vecteurs les plus courts possible et afin d'éviter la répétition. Ainsi, les consonnes [ʃ] et [ʒ] ne possèdent pas de trait coronal dans notre description.

Tableau 2b
Codification des phonèmes du corpus (consonnes)

API	ASCII	TRAITS							
					mode		lieu		
		V/C	voisé	sonant	constr.	occl.	labial	coronal	nasal
p	p	0	0	0	0	1	1	0	0
b	b	0	1	0	0	1	1	0	0
t	t	0	0	0	0	1	0	1	0
d	d	0	1	0	0	1	0	1	0
k	k	0	0	0	0	1	0	0	0
g	g	0	1	0	0	1	0	0	0
f	f	0	0	0	1	0	1	0	0
v	v	0	1	0	1	0	1	0	0
s	s	0	0	0	1	0	0	1	0
z	z	0	1	0	1	0	0	1	0
ʃ	S	0	0	0	1	0	0	0	0
ʒ	Z	0	1	0	1	0	0	0	0
l	l	0	1	1	0	1	0	1	0
r	r	0	1	1	1	1	0	0	0
m	m	0	1	1	0	1	1	0	1
n	n	0	1	1	0	1	0	1	1
ɲ	G	0	1	1	0	1	0	0	1
nulle	*	0	0	0	0	0	0	0	0

7. Expérimentation

7.1 L'entraînement du réseau

Nous avons entraîné le réseau avec les 900 noms les plus fréquents de notre corpus. Chaque mot encodé forme ce qu'on appelle une configuration d'entrée. Les configurations d'entrée obtenues pour les 900 noms ont été soumises au réseau 45 fois de façon non séquentielle. Le coefficient d'apprentissage était égal à 0,2 et le momentum[6] à 0,8. Le choix de ces paramètres a été réalisé après diverses tentatives : nous avons entraîné le réseau en variant le nombre de balayages, soit 10 000 puis 20 000, 30 000, jusqu'à 60 000; pour chacune

6 Le coefficient d'apprentissage et le momentum sont des paramètres du logiciel *Tlearn* qui servent à déterminer l'ajustement des poids des connexions entre les unités dans le but d'obtenir une diminution de l'erreur au cours de l'entraînement.

de ces valeurs, nous avons varié le coefficient d'apprentissage et le momentum de 0,1 à 1. Les valeurs retenues sont celles qui ont permis une meilleure performance du réseau. Dans un modèle d'apprentissage supervisé comme celui utilisé dans cette étude, l'erreur est représentée par la divergence entre la sortie attendue et la sortie obtenue. Elle est mesurée en termes de moyenne quadratique («RMS error»). La figure 1 montre la courbe RMS obtenue lors de l'entraînement du réseau avec les paramètres choisis.

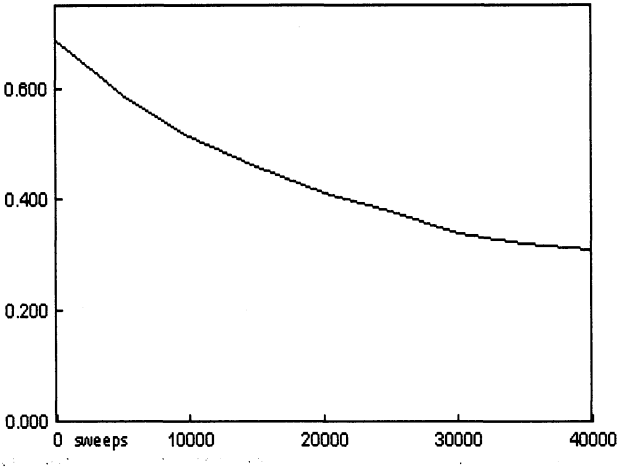

Figure 1 – Courbe RMS obtenue lors de l'entraînement du réseau.

7.2 La période de test[7]

Après la période d'entraînement, le réseau a été testé, d'abord avec les données d'entraînement (900 mots) et ensuite avec de nouvelles données, soit les 472 mots réservés pour évaluer sa capacité de généralisation. Les poids utilisés lors des tests sont ceux obtenus à la fin de l'entraînement. Nous avons utilisé la disposition «Output translation» de *Tlearn* pour identifier les sorties obtenues pour chaque configuration d'entrée. Cet outil convertit, à partir d'une table qu'on lui fournit, la configuration binaire des sorties en caractères ASCII (01 = F et 10 = M). Pour *Tlearn*, une unité de sortie est considérée comme activée (valeur 1) lorsque son niveau d'activation est plus grand que 0,5.

7 Le mot *test* est utilisé en connexionisme pour indiquer l'étape expérimentale ultérieure à l'apprentissage du réseau.

Les tableaux 3.1 et 3.2 montrent les résultats des tests réalisés avec les données de l'entraînement et les données nouvelles respectivement. Nous pouvons y observer que le réseau a été capable de classer correctement 96 % (n=900) des mots utilisés dans la période d'entraînement, mais qu'il n'a classé correctement que 56 % (n=472) des nouveaux mots.

Tableau 3.1
Résultats des tests – données de l'entraînement

	MOTS FÉMININS	MOTS MASCULINS	TOTAL	
EXACTS	422	444	866	96%
INEXACTS	26	8	34	4%
TOTAL	448	452	900	100%
	50%	50%		

Tableau 3.2
Résultats des tests – données nouvelles

	MOTS FÉMININS	MOTS MASCULINS	TOTAL	
EXACTS	144	121	265	56%
INEXACTS	107	100	207	44%
TOTAL	251	221	472	100%
	53%	47%		

8. Analyse des résultats

D'après les résultats présentés dans le tableau 3.1, le modèle connexioniste utilisé dans cette expérimentation a été capable d'assigner correctement le genre de 96 % des mots qui ont servi à son entraînement. Toutefois, selon les données présentées dans le tableau 3.2, le modèle n'a pas été capable de généraliser son apprentissage, ne classant correctement que 56 % des nouveaux mots. Pour pouvoir expliquer ces résultats, nous devons observer nos données plus en détail. D'abord, nous tenterons d'établir une relation entre les terminaisons des mots et le classement réalisé par le réseau, notamment en ce qui concerne les terminaisons considérées comme hautement prédictives. Ensuite, nous regarderons comment le réseau a effectué le classement, plus spécifiquement sur

quels traits communs aux noms il s'est basé pour les classer. Puisque les terminaisons suffixales et les mots composés ne forment qu'une partie minime de notre corpus, nous ne regarderons pour l'instant que les terminaisons phonétiques ayant une certaine valeur prédictive.

Le tableau 4 montre le nombre de mots masculins et féminins de notre corpus regroupés selon leur terminaison phonétique, ainsi que le pourcentage des mots classés correctement lors de la réalisation des tests. Le tableau montre aussi la valeur prédictive des terminaisons phonétiques selon Tucker, Lambert et Rigault 1977.

Tableau 4
Résultats selon la valeur prédictive des terminaisons

Terminaisons	Valeurs prédictives	Corpus d'Entraînement				Corpus de Test			
		Total	Mots fém	Mots masc	Bien classés	Total	Mots fém	Mots masc	Bien classés
œ	1.000M	3	0	3	100% M	0	0	0	N/A
ɑ̃	0.999M	26	1	25	100% M	10	0	10	60% M
ɛ̃	0.990M	30	1	29	93,1% M 100% F	15	0	15	46,7% M
ø	0.974M	8	0	8	100% M	2	0	2	50% M
o	0.972M	58	2	56	96,4% M 100% F	31	2	29	58,6% M 50% F
ɜ	0.942M	16	8	8	100% M 100% F	9	3	6	66,7% M 66,7% F
m	0.919M	19	10	9	100% M 100% F	10	2	8	87,5% M 0% F
ɛ	0.902M	21	3	18	100% M 100% F	15	0	15	46,7% M
f	0.890M	3	2	1	100% M 100% F	5	2	3	0% M 100% F
u	0.877M	14	0	14	100% M	3	2	1	100% M 100% F
a-ɑ	0.826M	28	9	19	100% M 88,9% F	11	3	8	63,6% M 66,6% F
r	0.750M	133	70	63	96,8% M 95,7% F	60	31	29	51,7% M 58% F
g	0.732M	0	0	0	N/A	3	2	1	0% M 0% F
y	0.716M	2	0	2	100% M	4	2	2	50% M 100% F
k	0.666M	21	9	12	100% M 100% F	10	8	2	0% M 50% F
b	0.651M	9	7	2	100% M 100% F	3	1	2	50% M 0% F
ɲ	0.390F	12	7	5	100% M 71,4% F	1	1	0	100% F
s	0.385F	37	33	4	75% M 90,1% F	37	33	4	25% M 60,6% F
d	0.381F	13	9	4	75% M 100% F	14	13	1	100% M 23% F
ʃ	0.340F	17	17	0	100% F	11	9	2	0% M 44,4% F
j	0.324F	34	22	12	100% M 100% F	12	4	8	62,5% M 50% F
n	0.315F	32	30	2	100% M 96,7 F	22	21	1	100% M 52,4% F
v	0.315F	7	5	2	100% M 100% F	5	5	0	100% F
ɔ	0.297F	74	2	54	100% M 75% F	52	21	31	51,6% M 76,2% F
i	0.246F	44	15	29	100% M 100% F	22	13	9	66,7% M 53,8% F
z	0.100F	16	16	0	100% F	8	8	0	62,5% F

Nous pouvons observer que notre corpus d'entraînement est constitué de 75,22 % (n=900) de mots ayant une terminaison à valeur prédictive du genre et que le réseau a bien appris à les classer. Comme nous l'avons vu à la section 5 de ce travail, Tucker, Lambert et Rigault 1977 ont démontré que les locuteurs natifs attribuent le genre à des noms qu'ils ne connaissent pas selon la proportion des noms masculins et féminins ayant la même terminaison que le nom inconnu. Dans notre simulation, nous nous attendions à ce que le réseau puisse généraliser son apprentissage et réussisse à classer les nouveaux mots selon la proportion des noms masculins et féminins du corpus. Autrement dit, nous nous attendions, par exemple, à ce que le réseau réussisse à classer comme masculins un minimum de 85 % des nouveaux mots finissant en [ɛ], puisque dans le corpus d'entraînement, 85,7 % des mots ayant cette terminaison sont masculins. Or, ce n'est pas ce qu'on constate en regardant les données du tableau 4 : le réseau réussit à classer correctement seulement 46,7 % des nouveaux mots masculins alors que leur proportion réelle s'élève à 100 %. Le classement des mots du corpus de test semble donc être le résultat du hasard plutôt que d'un apprentissage antérieur.

Pour bien comprendre ces résultats, nous devons regarder sur quelle base le modèle a réalisé la classification des noms. Pour ce faire, il faut vérifier le contenu des unités cachées. *Tlearn* nous permet de réaliser une analyse par groupements («cluster analysis») pour caractériser globalement la configuration des unités cachées. Le résultat de cette méthode d'analyse est une structure arborescente qui regroupe les données ayant des configurations d'unités cachées semblables (Bechtel et Abrahamsen 1993). La figure 2 montre un échantillon de la structure arborescente obtenue à partir du test avec les noms du corpus d'entraînement. Comme nous pouvons l'observer, les regroupements se font en considérant toutes les particularités possibles du nombre et du type des phonèmes. La terminaison ne semble être qu'un des aspects considérés.

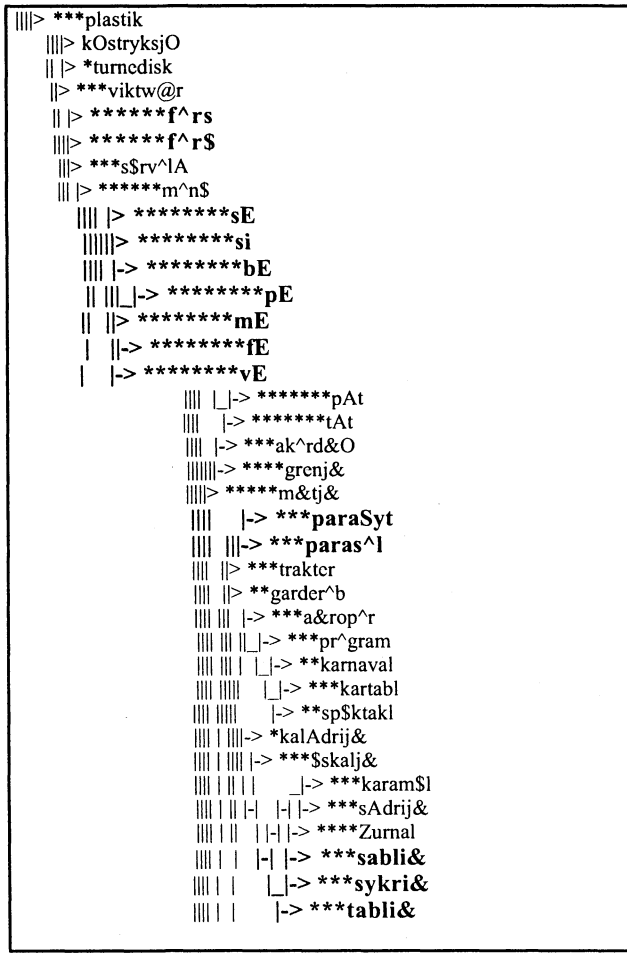

Figure 2 – Analyse par groupements des données de l'entraînement. Le système utilise toutes les informations possibles afin de parvenir à un groupement. Par exemple, [f^rs] et [f^r$] (*force* et *forêt*) sont visiblement classés selon les trois premiers phonèmes. Dans certains cas, ce type de groupement réussit (cf. [paraSyt] et [paras^l]). Les terminaisons semblent également être prises en compte (cf. [sabli&], [sykri&] et [tabli&]. Cependant, aucune réelle généralisation n'est effectuée, et souvent le système regroupe des terminaisons semblables qui ne diffèrent que par quelques traits comme [kartabl] et [sp$ktakl], où la terminaison ne diffère que par les traits de [b] et de [k].

En considérant ces résultats, nous avons décidé de réaliser deux autres expérimentations en utilisant la méthodologie décrite ci-dessus, mais en apportant des changements dans l'encodage des noms pour le premier cas et dans la taille du corpus pour le deuxième.

9. Expérimentation 2

Dans cette nouvelle simulation, nous avons encodé les noms de la même façon que Sokolik et Smith 1992. C'est-à-dire que nous avons utilisé un paradigme phonème/position : chaque phonème composant le nom est représenté par un vecteur de 35 bits, où le bit correspondant à la position du phonème dans notre table de conversion est activé. Nous avons procédé de cette façon pour diminuer la quantité de traits caractérisant les noms, dans le but de les distinguer plus amplement les uns des autres. Les résultats obtenus dans cette deuxième expérimentation sont sensiblement pareils à ceux de la première expérimentation réalisée. Le réseau apprend à classer correctement près de 100 % des mots présentés dans la phase d'entraînement, mais n'est capable de classer correctement que 52 % des mots inconnus. L'analyse par groupements réalisée avec les données de la période de test nous a permis d'observer encore une fois que le réseau se sert de toutes les caractéristiques du mot pour effectuer les regroupements.

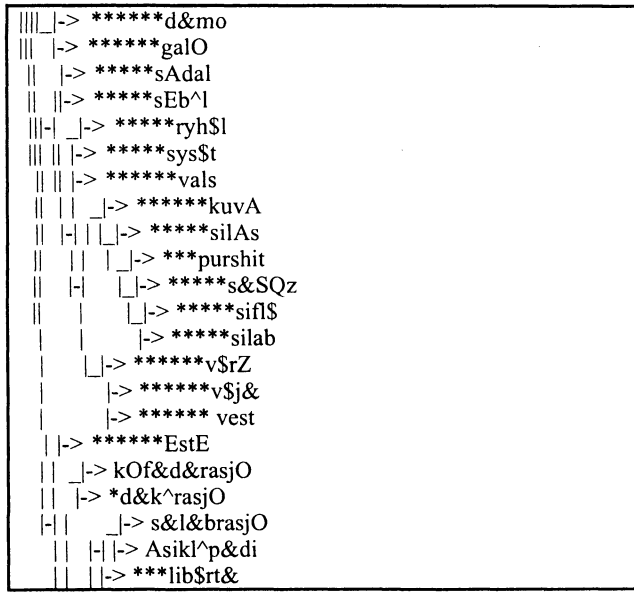

Figure 3 – Analyse par groupements des données du test de l'expérimentation 2. Comme pour la première expérimentation, le système regroupe les mots en tenant compte de toutes les informations possibles.

10. Expérimentation 3

Dans la présente simulation, nous avons utilisé comme corpus d'entraînement du réseau les 200 noms les plus fréquents de notre corpus original. Pour le corpus de test (les mots nouveaux), nous avons utilisé les 100 mots suivants. Le réseau a réussi à classer correctement 97 % des noms appartenant au corpus d'entraînement, ce qui correspond sensiblement aux mêmes taux de réussite que dans les expérimentations précédentes. Toutefois, dans cette troisième expérimentation, le réseau a été capable de classer correctement 71 % des noms du corpus de test. La figure 4 montre l'analyse par groupements pour le test réalisé avec les mots nouveaux : nous pouvons noter qu'il n'y a pas de différence par rapport aux regroupements effectués par le réseau dans les simulations précédentes. Il semble donc que l'amélioration de la performance du réseau dans le classement des mots nouveaux soit uniquement une conséquence du changement de corpus. En variant le nombre des mots du corpus d'entraînement et de test, nous avons fait varier les corrélations existantes entre les traits extraits par le réseau pour classer les noms et le genre de ces derniers. De toute évidence, il y a dans ce nouveau corpus une tendance à la convergence de ces deux types de classements qui n'existait pas dans le premier corpus.

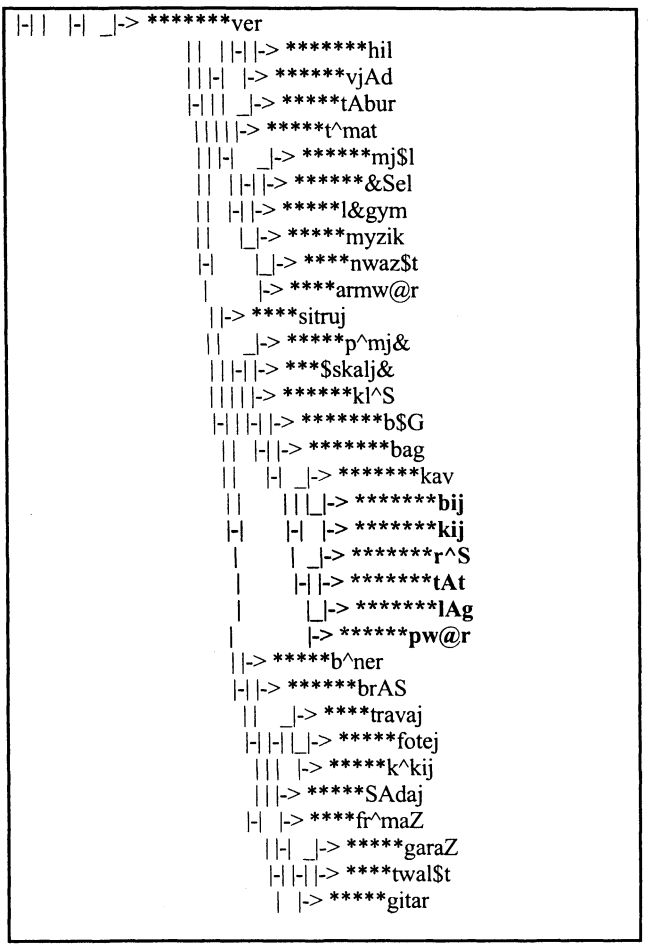

Figure 4 – Analyse par groupements des données du test pour l'expérimentation 3. Encore ici, le groupement se fait de manière identique aux deux premières expérimentations.

11. Conclusion

Nous avons vu que l'assignation du genre en français est un problème difficile à résoudre à l'aide d'un système connexioniste simple. Bien que les résultats obtenus par Sokolik et Smith 1992 peuvent de prime abord sembler concluants, nous avons démontré à l'aide des unités cachées que l'analyse connexioniste peut difficilement rendre compte du fonctionnement du langage.

Lors des expérimentations que nous avons faites, le système réussissait à classer correctement une très forte majorité des mots lors de la période d'entraînement, mais ce taux de réussite paraissait devenir aléatoire lors de la période de test. Au delà du fait que cette divergence semble démontrer l'inaptitude du système à généraliser, nous sommes amenés à mettre en doute le fonctionnement même du système et plus particulièrement le choix des unités de sortie et leur rôle dans l'apprentissage du système.

La préspécification des genres dans les unités de sortie est un choix qui ne reflète en rien les compétences linguistiques du locuteur, comme le soulignait Carrol 1995. Abandonner la spécification des unités de sortie permettrait de remédier à cette situation en plus d'empêcher le système d'agir comme «teacher». Comme nous l'avons vu, l'analyse par groupements montre bien que le système développe une catégorisation basée sur plusieurs facteurs (nombre de phonèmes, identification de la position des phonèmes, correspondance entre phonèmes, etc.), et ce tant au niveau de la période d'apprentissage qu'au niveau de la période de test. Il nous est donc permis de croire que même sans «teacher», le système serait amené à développer une classification. Demeure le problème de l'encodage des données.

Comme le mentionne Surridge 1993, l'assignation du genre en français est premièrement régie par des facteurs morphologiques, puis par des facteurs phonétiques. Nous croyons qu'un encodage morphophonologique pourrait peut-être amener des résultats plus adéquats.

Pour terminer, nous aimerions soulever une dernière hypothèse quant aux résultats présentés dans ce travail. L'être humain est constamment aux prises avec le temps. Quoi que nous fassions, le temps joue un rôle primordial et la perception du langage n'y échappe pas. Le découpage phonologique, syllabique ou morphologique s'effectue dans la durée temporelle. Lors de l'assignation du genre, les phonèmes ou les morphèmes qui indiquent le genre sont analysés linéairement. Or, le système connexioniste ne fait pas d'analyse linéaire. Le mot est présenté dans son entier aux unités d'entrée et l'analyse est effectuée sur l'ensemble de ces unités. Pour le système, il n'existe aucune différence d'importance entre le premier morphème ou le dernier. Nos résultats démontrent bien ce fait. La linéarité sera l'un des facteurs essentiels à prendre en compte lors des recherches ultérieures impliquant l'analyse phonologique ou morphologique du langage dans un cadre connexioniste.

Références

BARBAUD, P., C. DUCHARME et D. VALOIS 1982 «D'un usage particulier du genre en canadien-français : la féminisation des noms à initiale vocalique», *Revue canadienne de linguistique* 27-2 : 103-133.

BECHTEL, W. et A. ABRAHAMSEN 1993 *Le connexionisme et l'esprit – introduction au traitement parallèle par réseaux*, Paris, La Découverte.

CARROL, S. E. 1995 «The hidden dangers of computer modelling : remarks on Sokolik and Smith's connectionist learning model of French gender», *Second Language Research* 11-3 : 193-205.

DESROCHERS, A. 1986. «Genre grammatical et classification nominale», *Revue canadienne de psychologie* 40-3 : 224-250.

DESROCHERS, A., A. PAIVIO et S. DESROCHERS 1989 « L'effet de la fréquence d'usage des noms inanimés et de la valeur prédictive de leur terminaison sur l'identification du genre grammatical», *Revue canadienne de psychologie* 43-1 : 62-73.

ELMAN, J. L., E. BATES, M. H. JOHNSON, A. KARMILOFF-SMITH, D. PARISI et K. PLUNKETT 1996 *Rethinking Innateness : a Connectionist Perspective on Development*, Cambridge (Mass.), MIT Press.

FAUSETT, L.V. 1994 *Fundamentals of Neural Networks*, Upper Saddle River, Prentice-Hall.

HAYKIN, S. 1999 *Neural Networks: A comprehensive foundation*, Upper Saddle River, Prentice-Hall.

LACKS, B. 1996 *Langage et cognition*. Paris, Hermès.

MARCUS, G. F. 1998 « Can connectionism save constructivism?», *Cognition* 66 : 153-182.

MEHROTRA, K., C. K. MOHAN, et S. RANKA 2000 *Elements of Artificial Neural Networks*, Cambridge (Mass.), MIT Press.

PLUNKETT, K. et V. MARCHMAN 1991 « U-shaped learning and frequency effects in a multi-layered perceptron : implications for child language acquisition», *Cognition* 38 : 43-102.

PRÉFONTAINE, R-R. et G. PRÉFONTAINE 1968 *Vocabulaire ORAL des enfants de 5 à 8 ans au Canada français,* Montréal, Beauchemin.

RUMELHART, D.E. et J.L. MCCLELLAND 1986 «On learning the past tenses of English verbs», dans D.E. Rumelhart et J.L. McClelland, *Parallel distributed processing : Explorations in the microstructure of cognition,* volume 2, *Psychological and biological models*, Cambridge (Mass.), MIT Press.

SOKOLIK, M. E. et M. E. SMITH 1992 «Assignment of gender to French nouns in primary and secondary language : a connectionist model», *Second Language Research* 8 : 39-58

SURRIDGE M. E. 1985 «Le genre grammatical des composés en français», *Revue canadienne de linguistique* 30-3 : 247-271.

SURRIDGE M. E. 1986 «Genre grammatical et dérivation lexicale en français», *Revue canadienne de linguistique* 31-3 : 267-283.

SURRIDGE M. E. 1989a «Le facteur sémantique dans l'attribution du genre aux inanimés en français», *Revue canadienne de linguistique* 34-1 : 19-44.

SURRIDGE M. E. 1989b «Le genre grammatical en français fondamental : données de base pour l'enseignement et l'apprentissage», *Revue canadienne des langues vivantes* 45-4 : 665-674.

SURRIDGE M. E. 1993 «Gender Assignment in French : The hierarchy of rules and the chronology of acquisition», *IRAL* 31-2 : 77-95.

TUCKER, G. R., W. E. LAMBERT et A. A. RIGAULT 1977 *The French speaker's skill with grammatical gender : an example of rule-governed behavior*, La Haye et Paris, Mouton.

Réseaugraphie

NIX, A., D. MESSER, P. SMITH et N. DAVEY. *A connectionist account of Spanish gender harmony*, [En ligne], 1995. [http://www.cs.herts.ac.uk/~comrajn/paper8.html] (24 novembre 1999).