

UNE APPROCHE CONNEXIONNISTE POUR LA TRADUCTION AUTOMATIQUE DES TEXTES ARABE EN FRANÇAIS

K. REZEG¹, M. T. LASKRI²

1 Département d'Informatique - Université Mohamed KHIDER Biskra
BP 145, 07000 Biskra, Email rezeg_khaled@yahoo.fr
2 Département d'informatique - Université BADJI Mokhtar Annaba
BP 12, 23000 Annaba, Email : mtlaskri@wissal.dz , laskri@yahoo.com

RESUME

La traduction automatique des textes d'origine humaine est une application fort complexe appelée à appréhender les univers textuels ouverts, sans aucune contrainte quant à leur nature ou à leur diversité. Pour résoudre cette problématique, plusieurs tentatives ont été entamées ayant à chaque fois pour objectif l'obtention d'une meilleure qualité de traduction. Mais devant les diverses ambiguïtés de la langue naturelle, ce problème de traduction est loin d'être simple à résoudre. En effet, en absence de la maîtrise du contexte la plupart des phrases sont ambiguës. C'est pour cette raison, que les recherches actuelles dans ce domaine se sont orientées d'abord vers la maîtrise du sens véhiculé par les phrases avant de procéder au processus même de la traduction. En effet, traduire sans comprendre mène directement vers un échec. Pour ce faire, et dans le but d'augmenter la qualité de traduction, nous proposons une approche neuronale pour la génération des différents cas sémantiques relatifs aux différents composants de la phrase pour cerner d'abord le sens et de générer ensuite la traduction dans la langue cible. Ce qui nous a permis d'obtenir des résultats satisfaisants, comparativement à des travaux similaires utilisant d'autres techniques. Effectivement, Les techniques connexionnistes sont robustes, elles ont la réputation d'offrir des bonnes capacités, elles sont parallèles et elles s'appuient sur les notions de calcul réparti et de raisonnement distribué.

Dans notre approche, nous avons utilisé un apprentissage supervisé par un réseau simplement récurrent d'ELMAN (SRN) pour apprendre les techniques de traduction avec une performance similaire à celle des humains, ce qui constitue une première phase de notre système, celle de l'apprentissage. Une deuxième phase du système constitue la partie test et généralisation pour mesurer le degré de généralisation du réseau par rapport à ce qu'il appris durant la phase d'apprentissage.

Le réseau de neurones utilise, tient compte de la représentation approfondie en utilisant les cas sémantiques et la représentation surfacique en utilisant les éléments qui indiquent la forme de la phrase.

MOT CLES : Traduction automatique, Apprentissage, généralisation, réseau de neurone, cas sémantique.

ملخص :

الترجمة الآلية للنصوص تمثل تطبيقا غاية في التعقيد، تهدف لترجمة النصوص بكل الأشكال دون التقيد بطبيعتها وتنوعها. و من أجل الوصول لترجمة آلية ذات نوعية جيدة عدة أعمال جادة أنجزت، لكن غموض و صعوبة اللغة جعل من حل هذا الإشكال بعيد المنال. عموما تكون جل الجمل غامضة و للتحكم بشكل جيد في عملية الترجمة الأوتوماتكية باحثي هذا المجال ركزوا أعمالهم على فهم الجملة في لغتها الأصلية أولا ثم إيجاد ما يقابلها في اللغة المراد الترجمة إليها، لأنه بكل بساطة الترجمة بدون فهم تؤدي إلى فشل مؤكد.

من خلال عملنا هذا نهدف إلى إيجاد نظام ترجمة آلية ذات مستوى حسن بإستعمال نظام شبكات الخلايا العصبية الإصطناعية التي تمتاز بقدرة التعلم و التعميم، و قدرة مواجهة الخلل و القدرة على المعالجة المتوازية، معتمدين على نظرية حالات المعنى (نظرية فيلمور 1968) التي تسمح بتحديد معنى الجملة من خلال تحديد دور كل عنصر من عناصرها، ثم القيام بالترجمة في اللغة المقابلة مما سمح لنا الحصول على نتائج مقبولة مقارنة بتقنيات أخرى. و تم تجسيد نظامنا من خلال التعليم المراقب مستعملين الشبكة التراجعية البسيطة لإلمان لتعلم تقنيات الترجمة ب جودة مشابهة للترجمة العادية و هو ما يمثل المرحلة الأولى لنظامنا و التي تليها مرحلة ثانية تسمح بإستغلال ما تم إنجازه في المرحلة الأولى لإتمام مرحلة التعميم معتمدين على التمثيل المعمق للجملة مستعملين حالات المعنى و التمثيل الظاهري للجملة الذي تحدده مجموعة من العناصر التي تبين شكل الجملة الخارجي .

الكلمات المفاتيح : الترجمة الآلية، التعليم، التعميم، شبكة الخلايا العصبية الإصطناعية، حالات المعنى

1 INTRODUCTION

Comprendre une langue étrangère, traduire un texte, apprendre à parler sont des exemples de tâches effectuées par l'être humain d'une manière instantanée, alors que l'ordinateur le plus puissant en est totalement incapable. Pour étudier ce type de tâches on fait appel à l'approche cognitive. Cette dernière est pluridisciplinaire. En effet, on voit linguistes, informaticiens et psychologues essayer de travailler côte à côte. Elle est utilisée face aux problèmes complexes, mal ou peu spécifiés qui nécessitent dans leur traitement une masse d'informations énorme de nature différente. Pour leur traitement automatique une première approche représentée par les systèmes à base de règles est utilisée, mais elle a montré son inadéquation avec ce type de problème. Une autre approche à savoir l'approche neuronale est devenue ces dernières années, la plus appropriée à ce type de problème. Les techniques connexionnistes sont généralement associées à des mécanismes d'apprentissage et de généralisation, elles sont capables de généraliser les connaissances apprises à de nouveaux exemples non appris. Contrairement à un système symbolique, le système connexionniste se construit lui-même, ce qui représente son point fort et son utilisation dans les domaines mal ou peu spécifiés, sa robustesse et sa réputation d'offrir des bonnes capacités d'apprentissage et de généralisation. Il est parallèle et s'appuie sur les notions de calcul réparti et le raisonnement distribué.

Dans cet article nous allons faire un survol sur le traitement automatique du langage naturel d'une manière générale et sur la traduction automatique d'une manière particulière. Puis nous passons à la description de notre modèle basé sur une approche connexionniste de la traduction automatique des langues naturelles. Enfin nous présentons nos résultats expérimentaux et nous terminons par une conclusion en spécifiant nos résultats et en dégageant les perspectives ouvertes par le présent travail.

2 LE TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL (T.A.L.N.)

Le T.A.L.N est le domaine de l'informatique linguistique appliquée concernant le traitement du langage naturel par ordinateur. Les applications T.A.L.N tournent autour de la communication Homme-Machine et celles relatives à la compréhension automatique des textes [Arc-95].

Un système de Traitement Automatique du Langage Naturel est composé de trois modules essentiels à savoir [Las-95] [Mef-00]

- Le module de compréhension (analyseur) ;
- Le module d'interprétation ;
- Le module de génération

Vu la complexité des connaissances nécessaires à un système de compréhension automatique du langage naturel. On parle usuellement d'éléments morphologiques, syntaxiques, sémantiques, pragmatiques, contextuels,

connaissances sur le monde ou l'interlocuteur [Tai-97].

3 LA TRADUCTION AUTOMATIQUE

Depuis 50 ans, l'informatique a énormément évolué, mais aucune machine n'est aujourd'hui capable de traduire un texte avec une qualité acceptable. Les connaissances en linguistique nécessaires à la traduction sont encore insuffisantes [Gro-97].

La traduction automatique est une application fort complexe qui est appelée à appréhender les univers textuels ouverts, sans aucune contrainte quant à leur nature ou à leur diversité.

3.1 La traduction

Ce domaine implique la compréhension d'une langue, le transfert dans une autre langue et la génération dans cette dernière. La complexité de la traduction automatique est due à la langue qui est considérée comme un organisme vivant non un simple lexique, et le degré d'isomorphisme entre les langues varie considérablement [Nao-02]. Il s'agit d'un système d'expression et de communication commune à un groupe social, possédant une culture et une civilisation donnée. Les différentes langues ne réfèrent pas à un même espace de sens mais à plusieurs espaces, différents par leur culture et leur civilisation. Chaque langue a sa perception de la réalité décrite et son propre découpage de la réalité. Ces différents découpages font que la correspondance de langue à langue est généralement imparfaite, du point de vue sens, et bien souvent divergente de la structure de surface [Gen-94]. Tout texte écrit dans une langue naturelle comporte une part d'ambiguïté. On dit qu'un énoncé est ambigu s'il a plusieurs sens plausibles. Ils existent cinq types d'ambiguïté :

- L'ambiguïté lexicale ;
- L'ambiguïté de structure de surface ;
- L'ambiguïté de structure profonde ;
- L'ambiguïté sémantique ;
- L'ambiguïté Intrinsèque.

La traduction d'un texte n'est pas limitée à une transposition de langue à langue mais aussi culturelle. Donc toute prétention à une traduction totale et exacte est d'avance vouée à l'échec et que la traduction est approximative.

3.2 La traduction automatique

La traduction automatique est l'application de l'informatique à la traduction de textes en langage naturel. Les systèmes de la traduction automatique sont traités suivant plusieurs méthodes à savoir :

1. La méthode directe : chaque mot, syntagme de la langue

source, est **lié** directement à l'unité correspondante de la langue cible [Int-01].

2. La méthode par transfert : dans cette méthode nous devons passer par l'**analyse** qui décrit de manière linguistique le document source; puis nous appliquons les règles de **transfert** qui décrivent les équivalences linguistiques et structurelles entre les langues. Enfin nous passons à la **génération** qui utilise les données linguistiques de la langue cible et produit un document en langage naturel cible [Int-01].
3. La méthode inter langue par langage pivot : dans cette méthode il y a deux étapes :
 - Le texte source est *analysé* et *représenté* dans les termes conformes aux règles du langage pivot qui est un **langage neutre**, artificiel, ou commun à plusieurs langues;
 - la *génération* du texte en langue cible s'opère à partir de cette représentation "pivot"[Int-01].

3.3 Les logiciels de traduction

Un logiciel de traduction est une application informatique qui permet d'obtenir de façon automatique une traduction de tout type de textes (lettres, rapports, articles, sites web...) d'une langue (source) vers une autre langue (cible). Un logiciel de traduction est composé de :

4. Règles linguistiques : le moteur de traduction est le cœur technologique du logiciel de traduction. Les étapes de traduction automatique sont les suivantes :
 - analyse du texte source
 - application de règles linguistiques de transformation et de désambiguïsation
 - génération de la traduction

Plus ces règles sont fines plus la traduction obtenue est précise [Int-02].

5. **Dictionnaires** : les dictionnaires intégrés aux logiciels de traduction ne sont pas seulement une liste de mots ou d'expressions avec leur traduction. Chaque mot ou expression doit être défini avec des informations linguistiques (morphologie, sémantique, syntaxe...) dans la langue source puis dans la langue cible. Ces informations sont ensuite gérées par le moteur de traduction. **Plus les dictionnaires sont riches, plus la traduction obtenue est précise** [Int-02].

6. **L'interface** : l'interface des logiciels de traduction a un rôle important, puisque c'est elle qui facilite la réalisation et l'exploitation de la traduction. En effet, une fois que le texte est traduit l'utilisateur a besoin de le relire, de le réviser et de le personnaliser pour l'adapter à son style, à son activité, à ses besoins. Ce besoin de révision et d'adaptation dépend de l'utilisation finale de la traduction (compréhension, diffusion en interne, publication externe...) [Int-02].

4 DESCRIPTION DU SYSTEME

L'idée de base de notre modèle est la suivante : «Toute phrase a :

7. Une représentation interne ou bien approfondie qui désigne le sens ; cette représentation est assurée par un formalisme de représentation interne universel pouvant décrire n'importe quelle idée indépendamment de sa structure surfacique. Dans notre cas ce formalisme est basé sur la théorie des cas sémantiques.
8. Une représentation surfacique : le sens est indépendant de toute structure de surface d'une phrase mais cette dernière est très indispensable pour l'exprimer. Cette représentation indique la forme de la phrase qui est déterminée par un ensemble d'éléments comme par exemple : le genre du nom : féminin/masculin, le temps du verbe : Présent /passé / futur..... »

Notre objectif vise à développer un système de traduction automatique qui conserve le sens et donne une représentation surfacique similaire à celle de la phrase en entrée avec performance.

Pour atteindre cet objectif, nous avons mis en œuvre un Traducteur Automatique des Langues Naturelles par une Approche Connexionniste (T.A.L.N.A.C.) basé sur les cas sémantiques pour la génération du sens et en faire une génération dans une langue cible. Nous utilisons un apprentissage supervisé par un réseau simplement récurrent (SRN) pour apprendre les techniques de traduction avec une performance similaire à celle des humains. Notre système sert à traduire une phrase en langue source (arabe) à la langue cible (française) avec une approche neuronale.

4.1 Architecture du système

Le système «T.A.L.N.A.C» a l'architecture suivante :

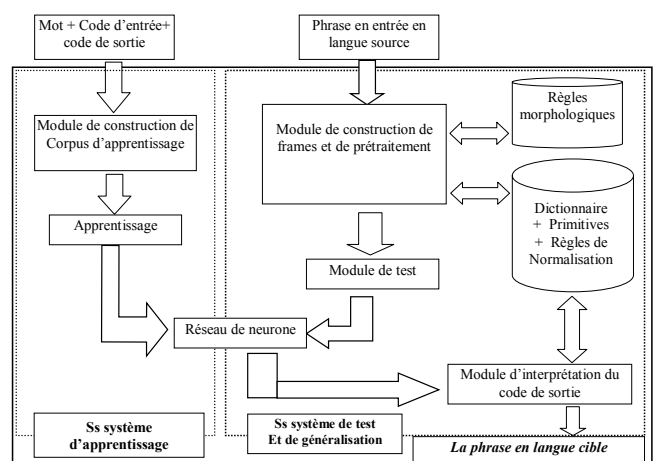


Figure 1 : Architecture du T.A.L.N.A.C

4.2 Topologie du réseau

Dans notre modèle nous avons utilisé un réseau simplement récurrent d'ELMAN qui a été utilisé pour les nombreuses

applications du traitement automatique du langage naturel [Tou-92], [Elm-90], [JW-90], [Berg-92], [Jod-93] avec un apprentissage supervisé en utilisant l'algorithme de rétropropagation et une représentation semi distribuée manuelle définie comme suit :

4.2.1 La couche d'entrée

La couche d'entrée comporte les neurones suivants :

a). Les neurones représentant les cas sémantiques : le tableau suivant contient le rôle et l'indice du neurone correspondant

Rôle	Indice
Action	1
Agent	2
Objet	3
Instrument	4
Manière	5
Patient	6
Lieu	7
Temps	8
Etat	9
Source	10
Destination	11
Fournisseur	12
Bénéficiaire	13
But	14

b). Les neurones représentant les primitives des verbes selon la classification de Shanck : cet ensemble de primitives sera utilisé dans l'opération de translation du verbe de la langue source (arabe) à la langue cible (française) (ces primitives seront utilisés dans la structure du dictionnaire).

Le tableau suivant présente l'indice du neurone et la primitive associée :

c). Les neurones représentant l'aspect surfacique de la phrase.

Le tableau suivant présente l'indice du neurone et l'aspect surfacique associé.

Indice	27	28	29	230	31	32	33
Aspect	Nombre	Genre	Défini	ماضي	حاضر	مضارع	أمر

d). Les neurones représentant la position du mot dans la phrase d'entrée :

Nous avons utilisé 08 positions

Indice	34	35	36	37	38	39	40	41
Position	Première	Deuxième	Troisième	Quatrième	Cinquième	Sixième	Septième	Huitième

e) Un neurone indiquant la forme de la phrase : Une phrase est soit passive (0) ou active (1).

Donc Cette couche contient 42 neurones.

4.2.2 La couche de sortie

Cette couche contient les neurones suivants :

a). Les neurones représentant la position du mot : On a utilisé 08 positions

Indice	1	2	3	4	5	6	7	8
Position	Première	Deuxième	Troisième	Quatrième	Cinquième	Sixième	Septième	Huitième

b). Les neurones représentant l'aspect surfacique de la phrase : le tableau suivant présente l'indice du neurone et l'aspect surfacique associé.

Indice	9	10	11	12	13	14	15
Aspect	Nombre	Genre	Défini	Passé	Présent	Futur	Impératif

Donc Cette couche contient 15 neurones.

Notre tâche consiste à lire une phrase mot par mot et le réseau décide la position et les éléments spécifiant l'aspect surfacique du mot.

Par exemple la phrase : يأكل الطفل كعكة sera traduite comme suit : **L'enfant mange un gâteau.**

Notre réseau possède la topologie suivante :

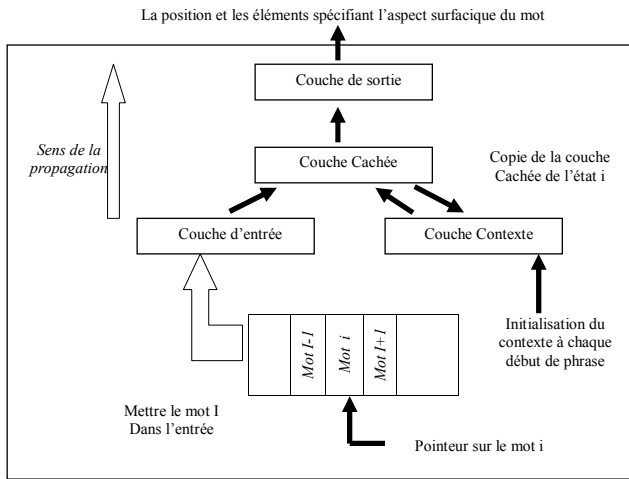


Figure 2 : Topologie et fonctionnement du réseau

Nous avons une fonction de propagation de type sigmoïde $f(x)=1/(1+e^{-x})$ (1)

4.3 Sous système d'apprentissage

Ce sous système sert à mettre à jour des poids des connexions à partir d'un ensemble de données d'entraînement jusqu'à l'obtention du comportement désiré. Il est composée de :

a). Le module de construction de corpus d'apprentissage : Ce dernier sert à assurer la cohérence du corpus d'apprentissage en appliquant un ensemble de règles par exemple : «Un argument ne peut porter qu'un cas et une phrase simple ne peut contenir un même cas qu'une fois au plus»

b). Le module d'apprentissage : l'apprentissage est effectué en utilisant les fonctions suivantes :

1- Formule de calcul de l'erreur en sortie: Permet de calculer l'erreur entre la sortie désirée (cible) et la sortie calculée.

$$Erreur = \sqrt{\sum_{i=1}^n (S_i - Y_i)^2} \quad (2)$$

N :nombre de neurone en sortie

2- Formule de calcul de l'erreur commise sur le neurone i :

$$a_i = 2*(S_i - Y_i) f'(I_i) \dots\dots (3)$$

i = 1..nombre de neurone de la couche de sortie

3- Formule d'ajustement des valeurs des poids : Permet d'ajuster les valeurs des poids synaptiques entre les neurones.

$$\Delta W_{ij} = \eta * a_i * S_j \quad (4)$$

η : pas d'apprentissage

a_i : Erreur commise sur le neurone i

S_j : Sortie du neurone j

4-Formule de calcul de l'erreur commise de la couche de cachée :

$$a_i = \sum_{j=1}^n w_{ij} * a_j * f(I_i) \dots \dots (5)$$

j = 1..nombre de neurone de la couche de sortie

i = 1..nombre de neurone de la couche cachée

b.2. Les paramètres d'apprentissage : Pour effectuer la phase d'apprentissage nous avons utilisé un ensemble de paramètres qui seront choisis d'une manière expérimentale. Ces paramètres sont :

1. Pas d'apprentissage $\eta = 0.6$;
2. Nombre de neurones de la couche cachée = 25 ;
3. L'erreur désirée : Elle est variable, Tant qu'il y a stagnation du taux de reconnaissance sur le corpus d'apprentissage nous devons la diminuer.

4.4 Sous système de test ou de généralisation

A travers ce sous système nous évaluons notre réseau qui dépend de sa capacité de généralisation de nouvelles phrases et nous enrichissons notre corpus d'apprentissage. Il est composé des modules suivants :

a). Module de construction de Frames et de pré traitement : ce module sert à :

1. introduire la phrase sous forme de frame dont les slots ayant les noms des cas sémantiques sont à remplir par les différents constituants de la phrase ;
2. effectuer un ensemble de traitement pour déduire le code d'entrée d'une manière automatique.

b). Module de test : ce module sert à activer le réseau pour trouver le code de sortie en exploitant le code d'entrée.

c). Module d'interprétation : ce module sert à interpréter les résultats du module de test. L'interprétation est effectuée à travers :

1. La recherche de la racine du mot : Pour trouver la racine du mot nous appliquons un ensemble d'opérations (enlèvement de préfixe, enlèvement de suffixe,...). Ce dernier doit exister dans le dictionnaire pour faire la translation dans la langue cible.
2. La présentation de la phrase en langue cible : cette opération représente le résultat final de la traduction. Elle est assurée à travers l'application des règles propres à la langue cible sur des informations tirées du dictionnaire.

4.5 Le dictionnaire

Pour assurer une translation des mots d'une manière correcte (acceptable) on doit utiliser un dictionnaire basé sur le principe des racines trilitères dont la structure nous donne la possibilité de bien choisir l'équivalent convenable dans la langue cible. Notre dictionnaire a la structure suivante :

1. Racine du mot : ce champ doit contenir la racine du mot ou bien du verbe dans la langue source.
2. Le schème du mot : ce champ doit contenir le schème du mot ou bien du verbe dans la langue source.
3. Le mot en langue source : ce champ doit contenir le mot ou bien le verbe dans la langue source.
4. Le rôle sémantique : ce champ a un rôle important dans la validité de l'opération de translation du mot de la langue source à la langue cible, mais comme notre travail concerne les phrases qui contiennent un seul verbe nous remplaçons le rôle sémantique pour les mots dont le rôle sémantique est une action par leur primitive.
5. Le type : ce champ contient le type du mot en langue source par exemple : Verbe, nom féminin, nom masculin,...
6. Groupe du verbe : indique le groupe de verbe en langue cible, cette information est très importante pour bien conjuguer le verbe par exemple : premier groupe / deuxième groupe ;
7. Auxiliaire du verbe : indique l'auxiliaire du verbe en langue cible, cette information est très importante pour bien conjuguer le verbe par exemple : Être / Avoir ;
8. L'équivalent : ce champ contient le mot équivalent du mot en langue source dans la langue cible.
9. Le pluriel de l'équivalent : ce champ contient le pluriel du mot équivalent du mot en langue source dans la langue cible.

5 RESULTATS EXPERIMENTAUX

Dans un système neuronal les résultats expérimentaux représentent un élément primordial pour tester sa validité. Pour effectuer le test, nous avons procédé comme suit :

1. construction de notre propre corpus : Cette phase représente la plus lente phase, elle nécessite une coordination entre linguistes et informaticiens pour arriver à former un corpus complet ;
2. adaptation du corpus d'apprentissage exploité dans le système de la génération d'une représentation du sens en langue arabe basée sur les cas sémantiques [Mef-00]
3. le lancement de l'apprentissage : Pour notre cas cette phase nécessite peu de temps.
4. Test et généralisation : Durant cette phase nous avons utilisé des corpus de généralisation de différentes natures (02 mots, 03 mots,...).

5.1 Résultats des test de l'apprentissage

Les résultats de test du processus d'apprentissage sont présentés dans le tableau suivant dont les colonnes sont :

- **Erreur Favoris** : seuil d'erreur d'apprentissage permit ;
- **Erreur de corpus** : erreur d'apprentissage calculée
- **Temps en Seconde** : temps nécessaire pour compléter l'apprentissage ;
- Nombre de chargement de corpus (NB) : Indique le nombre de fois de chargement des corpus d'apprentissage pour compléter l'apprentissage.
- **Taux d'apprentissage** : représente le résultat d'apprentissage.

Dans notre cas nous voyons que le facteur temps est liée à la vitesse de la machine, pour cette raison nous ajoutons un autre paramètre d'évaluation de notre apprentissage à savoir **le nombre de fois de chargement du corpus d'apprentissage**. (L'étape de Chargement des phrases à partir du fichier de corpus d'apprentissage est correspondante à l'étape 2 de l'algorithme de la rétro-propagation (Présentation d'un couple (entrée, sortie désirée)). Les résultats obtenus pour notre propre corpus sont indiqués dans le tableau 1.

5.2 Interprétation des résultats d'apprentissage

D'après les résultats de processus d'apprentissage des différents corpus utilisés nous pouvons tirer les conclusions suivantes :

1. Vu que les sorties de notre système sont de natures différentes (position, nombre, genre, défini, temps), nous jugeons que les taux d'apprentissage sont acceptables pour tous les types de corpus d'apprentissage (Plus de 90 %).
2. Notre réseau converge généralement pour des intervalles de temps acceptables.

5.3 Résultats des test de la généralisation

Dans le but de tester les résultats de la phase de l'apprentissage, nous avons effectué une série de test avec des corpus de différentes tailles (02 mots, 03 mots, 04 mots, de tailles variables) dont le nombre de phrases de chaque corpus égale 60 phrases sauf dans le cas de phrases de corpus de tailles variantes nous avons utilisé le corpus du système du réseau simplement récurrent pour la génération d'une représentation du sens d'une phrase en langue arabe basée sur les cas sémantiques» [Mef-00] (80 phrases).

Les résultats obtenus dans les tests de la généralisation sont indiqués dans le tableau 2.

Tableau 1 : Tableau récapitulatif du processus d'apprentissage en variant la taille de corpus

Erreur Favoris	Corpus de 60 phrases				Corpus de 70 phrases				Corpus de 80 phrases				Corpus de plus 90 phrases			
	Erreur	Temps	NB	Taux	Erreur	Temps	NB	Taux	Erreur	Temps	NB	Taux	Erreur	Temps	NB	Taux
0.20	0.1948	1.71	12	89.34	0.1934	1.98	12	87.23	0.1989	2.69	14	88.88	0.1979	3.46	14	89.73
0.19	0.1897	2.30	16	89.34	0.1889	3.24	19	91.06	0.1896	3.13	16	88.51	0.1888	3.85	16	88.56
0.18	0.1799	2.42	17	87.30	0.1799	3.74	22	86.38	0.1791	4.12	21	90.00	0.1797	4.67	19	88.56
0.17	0.1613	3.51	25	89.84	0.1657	3.51	21	91.48	0.1671	4.83	25	90.74	0.1685	4.39	18	88.85
0.16	0.1560	2.86	20	91.37	0.1589	4.62	27	91.91	0.1591	5.66	29	91.74	0.1573	5.60	23	82.99
0.15	0.1468	3.13	22	90.86	0.1474	2.85	17	88.51	0.1489	6.81	35	91.85	0.1499	7.31	30	90.61
0.14	0.1399	4.12	29	90.86	0.1338	4.94	29	91.06	0.1399	6.59	34	88.14	0.1390	9.55	29	90.90
0.13	0.1294	6.59	46	92.38	0.1297	6.43	38	90.21	0.1294	7.85	40	91.48	0.1295	11.04	45	91.20
0.12	0.1138	5.71	40	92.89	0.1198	9.17	54	91.06	0.1199	9.62	49	92.22	0.1199	17.03	69	93.54
0.11	0.1097	5.88	41	91.37	0.1099	5.28	31	91.91	0.1098	15.98	82	94.07	0.1100	22.14	90	93.54
0.10	0.0960	12.47	87	94.92	0.0938	10.54	62	92.34	0.0996	24.06	123	93.70	0.1000	21.42	87	94.13
0.09	0.0897	11.04	77	95.43	0.870	6.98	41	91.06	0.0898	25.59	131	94.44	0.0895	53.93	219	94.72
0.08	0.0755	9.34	65	92.38	0.0797	11.97	70	92.34	0.0785	40.48	207	97.40	0.0798	87.39	354	95.30
0.07	0.0687	9.28	65	93.90	0.0684	20.05	117	94.89	0.0699	35.92	184	95.92	0.0699	84.04	342	95.89
0.06	0.0596	11.20	78	93.90	0.0596	18.07	106	95.74	0.0599	75.58	387	94.81	0.0580	58.11	237	96.18
0.05	0.0497	67.56	469	97.46	0.0497	118.97	695	96.59	0.0453	71.79	368	97.03	0.0495	87.93	359	95.30

Tableau 2 : Tableau des résultats des tests de la généralisation en variant les tailles des phrases

Nature du corpus	Phrases à 02 mots	Phrases à 03 mots	Phrases à 04 mots	Phrases de taille variable
Taux de reconnaissances	96.66	93.33	88.33	87.50

5.4 Interprétation des résultats de généralisation

D'une manière générale les résultats de la phase de généralisation sont satisfaisants de telle manière que les positions des différents mots de la phrase traduite sont bien calculées et dans la majorité des cas nous avons obtenu des représentations surfaciques justes.

6 IMPLEMENTATION

Dans la phase d'implémentation de notre modèle nous avons utilisés exploités le langage de programmation DELPHI 05 et son module de bases de données (paradox) sous Windows. Notre modèle est exploité à travers une interface principale assurant les fonctionnalités de base comme l'apprentissage et la généralisation ainsi que les autres fonctionnalités secondaires comme la mise à jour du dictionnaire, le help,...

6.1 Fiche d'apprentissage

Après la construction du corpus d'apprentissage, l'opération d'apprentissage qui consiste à l'application de

l'algorithme de la rétropropagation sur les données de corpus d'apprentissage est effectuée à travers la fiche suivante :

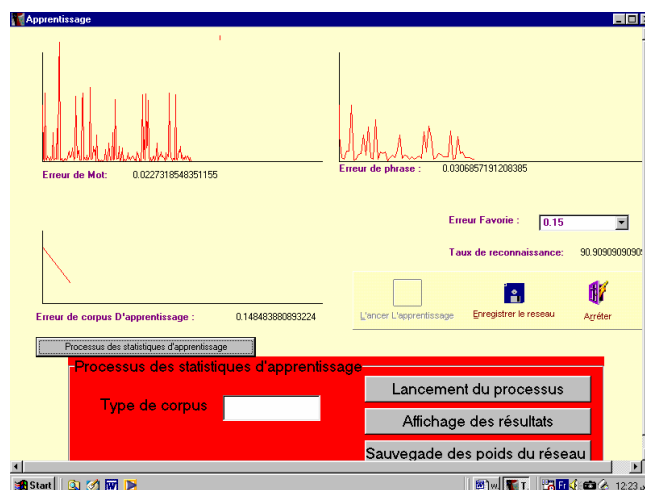


Figure 3 : Fiche d'apprentissage

Dans notre travail nous avons exploité de l'algorithme de la rétropropagation sur les données de corpus d'apprentissage comme suit :

Pour chaque erreur (de 0.2 à 0.05) faire

1. Initialisation du réseau ;
2. Chargement des phrases à partir du fichier de corpus d'apprentissage ;

3. Application de l'algorithme de la rétropropagation sur les données de corpus
4. Calcul de taux de reconnaissance
5. Si ((l'erreur de corpus est supérieur à l'erreur (celle favoris)) ou (le nombre de chargement a partir du fichier du corpus <=5000)) aller à 2.

6.2 Fiche de la généralisation (traduction des phrases)

A travers cette fiche nous effectuons l'introduction des constituants de la phrase chacun dans le slot du cas sémantique correspondant, la traduction de la phrase et l'améliorer la qualité de la traduction. Les résultats peuvent être affichés sous forme de texte comme il est indiqué dans la figure 05.

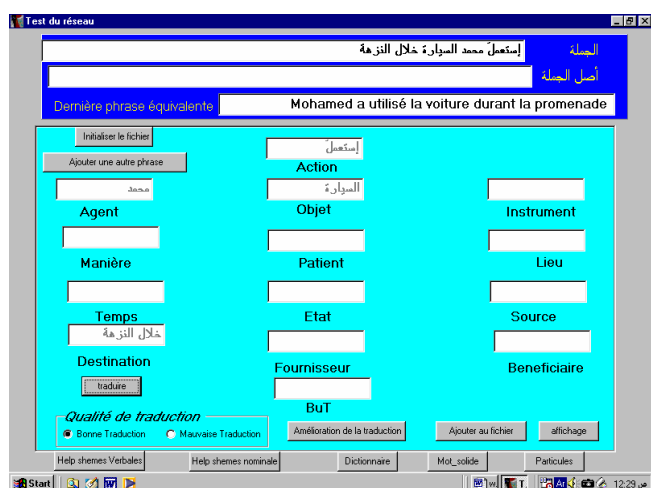


Figure 4 : Fiche de généralisation

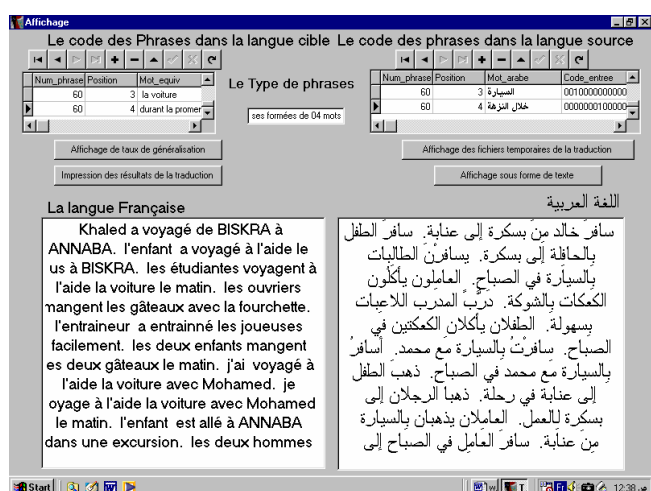


Figure 5 : Fiche d'affichage des résultats sous forme de texte

7 CONCLUSION

A travers cet article nous avons proposé un système de traduction automatique basé sur les cas sémantiques en utilisant un apprentissage supervisé par un réseau simplement récurrent (SRN). Nous pouvons dire que la réalisation de système est fortement liée aux points suivants :

1. Il faut bien choisir un corpus d'apprentissage complet qui permet d'avoir un réseau adéquat ;
2. Il faut arriver à un taux de reconnaissance soit dans la phase d'apprentissage ou dans la phase de généralisation très élevé, dans le cas contraire il faut faire recours soit aux éléments surfaciques, soit au corpus d'apprentissage.

Les perspectives de notre travail se résument en la réalisation d'un système de traduction automatique universel en développant :

1. Un réseau de neurone universel ou nous prenons les éléments de l'aspect surfacique comme paramètres ;
2. Des modules de pré traitement et d'interprétation propres aux langues ;
3. Un module de transformation des phrases complexes (qui contiennent plus qu'un verbe) à des phrases simples.

REFERENCES

- [1] [Arc-95] Dominique ARCHAMBAULT, «Proposition de réseaux neuromimétiques par le traitement du langage naturel» Thèse de Doctorat 3eme Cycle 1995
- [2] [Beg-92] Geroge BERG , «A connexionist parser with recursive sentence structure and lexical disambiguation» AAAI-92 American Association for artificial intelligence
- [3] [Elm-90] Jeffry ELMAN , «Finding Structure in Time » Cognitive science 14/179-211 1990
- [4] [Gen-94] G.GENELEX, «Rapport sur la couche sémantique. Projet Eureka Genelex. Copyright ASSTRIL, GSI-Erli , IBM France, Sema Group» 1994
- [5] [Gro-97] M.GROSS, «La traduction automatique, dans les langues du monde », dossier hors série pp. 126-130 » 1997
- [6] [Int-01] Traduction - Terminotique - Lexicographie sur le site de l'Observatoire Wallon de l'Info route & du Traitement Informatique des Langues <http://www.owil.org/traduc.htm#TA>
- [7] [Int-02] Livre Blanc Traduction, <http://www.softissimo.com/products/infotrad.htm>
- [8] [Jod-93] Jean François JODUIN , «Réseaux de neurones et traitement de langage naturel : Etude des réseaux de neurones récurrents et de leurs représentations», Thèse de Doctorat de l'université PARIS XI ORSAY 1993

- [9] [Las-95] Mohamed Tayeb LASKRI et R. MAHDJOUBI, «Traitement automatique de la langue arabe en vue d'une traduction automatique des textes vers la langue française». Proc. 3èmes Journées Internationales de traitement de données textuelles **JADT'95**. Rome, Italie, 11-13 Décembre 1995.
- [10] [Mef-00] Karima MEFTOUH, «Un réseau simplement récurrent pour la génération d'une représentation interne du sens en langue arabe basée sur les cas sémantiques» Mémoire de Magister, Université de Annaba 2000
- [11] [Nao-02] NAOUI Oum El Kheir, «Analyse automatique avec d'un groupe nominal (application à la traduction automatique)» Mémoire de Magister, Université Es-senia , Oran 2002
- [12] [Tou-92] Clause TOUZET, «Les réseaux de Neurones artificiels, Introduction au connexionnisme». Cours, Exercices et travaux pratiques Juillet 1992
- [13] [Tai-97] Nacera TAIBI Née MELLITI, «Contribution à l'étude du traitement automatique des erreurs dans un texte écrit en arabe» Mémoire de magister ENSLSH 1997