Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Évolution phonétique des langues et réseaux de neurones : travaux préliminaires

Clémentine Fourrier
Inria, 2 rue Simone Iff, 75012 Paris, France
clementine.fourrier@inria.fr

RÉSUMÉ

La prédiction de cognats est une tâche clef de la linguistique historique et présente de nombreuses similitudes avec les tâches de traduction automatique. Cependant, alors que cette seconde discipline a vu fleurir l'utilisation de méthodes neuronales, celles-ci restent largement absentes des outils utilisés en linguistique historique. Dans ce papier, nous étudions donc la performance des méthodes neuronales utilisées en traduction (les réseaux encodeur-décodeur) pour la tâche de prédiction de cognats. Nous nous intéressons notamment aux types de données utilisables pour cet apprentissage et comparons les résultats obtenus, sur différents types de données, entre des méthodes statistiques et des méthodes neuronales. Nous montrons que l'apprentissage de correspondances phonétiques n'est possible que sur des paires de cognats, et que les méthodes statistiques et neuronales semblent avoir des forces et faiblesses complémentaires quant à ce qu'elles apprennent des données.

ABSTRACT

Sound change and neural networks: preliminary experiments

Cognate prediction is a key task in historical linguistics that presents a number of similarities with machine translation. However, although neural methods are now widespread in machine translation, they are still largely unused in historical linguistics. In this paper, we study the performance of neural methods (more specifically encoder-decoder networks) for the task of cognate prediction. We focus in particular on the types of data that can be used for this task, and compare the performance of statistical and neural methods. We show that sound correspondances can only be learned using cognate datasets, and that statistical and neural methods seem to have complementary strengths and weaknesses regarding what they learn about the data.

MOTS-CLÉS : Changements phonétiques, réseaux de neurones, prédiction de cognats, linguistique historique.

KEYWORDS: Regular sound changes, neural networks, cognate prediction, historical linguistics.

1 Contexte

Osthoff *et al.* (2014) furent les premiers à formaliser la notion de régularité des changements phonétiques, pierre angulaire de la méthode comparative sur laquelle reposent toutes les études réalisées depuis en phonétique historique. Leur observation empirique était la suivante : une transformation diachronique d'un phone en un autre phone, dans un contexte donné, est régulière et sans exception ; à ce titre, tous les mots contenant ce même phone dans ce même contexte vont eux aussi subir cette transformation phonétique. Cette observation fut rendue possible par la comparaison des représenta-

Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 110–122.
HAL : hal-02786192.

tions phonétisées d'ensembles de mots appelés cognats [1], qui permit d'identifier des motifs récurrents dans les correspondances phonétiques.

L'utilisation itérative de cette méthode pour, successivement, identifier de nouvelles règles de changement phonétique en comparant des cognats, puis trouver des cognats inédits grâce aux règles nouvellement identifiées, constitue le cœur de la phonétique historique, et permet de définir deux de ses tâches principales : l'identification de cognats et la prédiction de changements phonétiques.

Avec l'émergence des méthodes informatiques, de nouvelles façons de traiter ces tâches, s'appuyant sur le traitement massif de nombreuses données, virent le jour. Une grande majorité d'entre elles reposent sur des comparaisons automatiques de lexèmes, qui combinent des méthodes d'alignement phonétique, soit avec des calculs de distances de type Levenshtein ou Turchin, soit avec des méthodes d'agrégation (List *et al.* (2017) font une comparaison de ces différents méthodes). Plus récemment, certains travaux commencèrent à capitaliser sur les fortes similitudes entre les tâches précédemment décrites et la traduction automatique : les deux visent à apprendre le passage d'une séquence d'éléments ordonnés à une autre séquence d'éléments ordonnés (problème « *many to many* »). Dans cette optique, Beinborn *et al.* (2013) utilisent Moses, logiciel appliquant des méthodes de traduction statistique (Koehn *et al.*, 2007), pour de la prédiction de cognats (avec une définition large de cognat).

Cependant, le domaine de la traduction automatique s'est maintenant enrichi de méthodes neuronales, et, au vu de leur prééminence dans la discipline, on est en droit de se demander pourquoi ces méthodes n'ont quasiment jamais été appliquées à l'heure actuelle pour les problématiques de linguistique historique, notamment en ce qui concerne le modèle de référence de la traduction : l'encodeur-décodeur. A notre connaissance, le seul travail de recherche réalisé pour l'instant sur l'application de cette architecture à ces problématiques est le rapport de mémoire de Dekker (2018), dont les conclusions sont peu reluisantes : pour la tâche de prédiction de cognats, un simple perceptron surpasse les réseaux récurrents de type encodeur-décodeur (en utilisant 1016 formes phonétisées appartenant à une centaine de langues).

Ces travaux ont donc soulevé la problématique suivante : est-il possible d'utiliser des réseaux de neurones de type encodeur-décodeur pour apprendre des correspondances phonétiques, et si oui, sur quel type de données ?

Nous tentons de répondre à ces questions, avec quelques expériences préliminaires. En premier lieu, nous étudions les types de données sur lesquelles il est possible d'apprendre des correspondances phonétiques, et notamment s'il est possible de pallier le faible nombre de données de cognats en utilisant des lexiques bilingues. La seconde expérience s'intéresse à la pertinence de l'utilisation et au paramétrage de réseaux de neurones pour cette tâche. Enfin, les dernières expériences visent à étendre, de façon préliminaire, ces résultats à des données réelles.

1. Des cognats sont des mots de langues différentes, descendants d'un même « mot-ancêtre » commun, celui-ci appartenant à une langue parente de toutes les langues concernées. Ces mots ont donc vécu toutes les évolutions phonétiques de leurs langues respectives. Par exemple, les mots polonais *być* 'être', tchèque *být* 'id.' et lituanien *būti* 'id.' sont tous cognats, et descendent d'un ancêtre commun en proto-balto-slave.

2 Cadre expérimental

2.1 Tâche à résoudre

Dans cet article, nous désirons étudier si un modèle neuronal peut apprendre des correspondances phonétiques. Nous définissons à cet effet la tâche d'intérêt comme étant la « traduction » de cognats phonétisés d'une langue vers une autre.

Cette tâche n'est pas triviale, et ce pour plusieurs raisons. Déjà, les données de cognats sont rares. Les jeux de cognats font en général de 100 à 200 mots, ce qui est très peu pour apprendre avec des méthodes neuronales sans sur-apprentissage (ces méthodes comprenant elles-même plusieurs centaines de poids mathématiques à ajuster). Ensuite, c'est une tâche qui peut présenter des ambiguïtés importantes, selon la direction de prédiction étudiée. Si l'on va d'une langue mère vers sa langue fille, les changements phonétiques sont strictement réguliers, et à un mot de la langue mère est associé un seul mot de la langue fille, sans ambiguïtés. Mais la transformation inverse n'est pas évidente : un mot de la langue fille pourrait, formellement, descendre de plusieurs mots de la langue mère [2]. Pour cette raison, nous demanderons à nos modèles de produire plusieurs prédictions, de une à trois, et nous comparerons l'impact du nombre de réponse prédites sur la précision. Enfin, les données qui sont accessibles sont généralement bruitées, et à ce titre, présentent un défi supplémentaire : en apprendre les correspondances phonétiques sans en apprendre le bruit.

2.2 Modèle étudié : modèle neuronal (MEDeA)

L'architecture neuronale que nous utilisons est un des modèles de référence en traduction automatique : l'encodeur-décodeur (Sutskever *et al.*, 2014) avec attention (Bahdanau *et al.*, 2015; Luong *et al.*, 2015). Ce type de modèle transforme l'entrée (une séquence ordonnée, de phones dans notre cas, de mots en traduction automatique) en une représentation intermédiaire vectorielle (« *hidden representation* ») grâce à l'encodeur, une succession de réseaux de neurones récurrents. Cette représentation intermédiaire est ensuite lue par le décodeur, qui prédit chaque phone (resp. mot) de la séquence finale successivement (dans la langue de sortie) en fonction de l'enchaînement des phones (resp. mots) précédemment prédits. Pour cet article, nous utilisons notre implémentation de cet algorithme, MEDeA (*Multiway Encoder Decoder Architecture*, en Python et PyTorch), qui utilise un encodeur différent pour chaque langue d'entrée, et un décodeur indépendant, avec sa propre attention, et différent pour chaque langue de sortie. Pour contraindre la représentation intermédiaire à un seul espace, le réseau apprend sur toutes les paires de langues possibles (y compris d'une langue vers elle-même). Notre architecture permet de passer d'un grand nombre de langues à un grand nombre de langues identiques ou non.

2. Par exemple, considérons un son [son1], dans la langue mère, qui évolue régulièrement en [son2] dans sa langue fille, tandis que [son2] dans la langue mère reste [son2] dans la langue fille. Pour passer de la langue mère à la langue fille, la situation ne présente pas d'ambiguïté : [son1] devient [son2], et [son2] devient [son2]. Par contre, à l'inverse, si on a [son2] dans la langue fille, il peut correspondre à [son1] comme à [son2] dans la langue mère.
Prenons un exemple concret. Dans certains cas, le [b] latin évolue en un [v] en italien, comme pour [kaˈbal.lus], 'cheval' qui devient en italien [kaˈvaːlo] 'id.'. Cependant, le son [v] latin reste un [v] en italien, comme dans [vita] 'vie', identique dans les deux langues. Un [v] italien correspond-il à un [v] ou à un [b] initial en latin ? Remonter de la langue fille à la langue mère présente ici une ambiguïté.

2.3 Modèle de référence : modèle statistique (Moses)

Moses est le modèle statistique de référence en traduction automatique statistique (Koehn *et al.*, 2007). Il fonctionne en deux étapes. Pour la première, l'entraînement, les données bilingues sont tokenisées et alignées avec l'aide de GIZA++ (Och & Ney, 2003), puis servent à l'apprentissage d'un modèle de langue tri-gramme de la langue de sortie, d'un tableau de correspondances entre les phones de la langue d'entrée et de la langue de sortie, et un modèle de réorganisation pour gérer l'ordre des phones. Lors de la seconde étape, la mise au point (« *fine tuning* »), les poids respectifs de chacun de ces modèles dans le système global sont ajustés grâce à un jeu de données de développement. Il est à noter qu'une différence notable entre le modèle neuronal et le modèle statistique est que le second ne peut apprendre que la traduction d'une langue vers une autre, là où le modèle neuronal apprend la correspondance de plusieurs langues en même temps.

2.4 Métrique d'évaluation : BLEU

Notre métrique d'évaluation est BLEU (Papineni *et al.*, 2002), plus spécifiquement l'implémentation SacreBLEU (Post, 2018). Le score BLEU calcule le pourcentage d'éléments communs (mots en traduction automatique, phones dans notre cas) entre la séquence de départ et celle d'arrivée, des unigrammes aux quadrigrammes d'éléments. La critique usuelle de cette métrique est qu'elle tend à mal noter des traductions correctes mais non présentes dans le jeu de référence ; cette critique ne s'applique pas à nos expériences, dans la mesure où il n'existe qu'une seule « traduction » possible d'un cognat en son équivalent dans une autre langue.

3 En l'absence de jeux de cognats, les lexiques bilingues peuvent-ils être utilisés pour étudier l'évolution phonétique des langues ?

3.1 Données

Notre tâche est définie comme l'apprentissage, à partir de paires de cognats, de régularités phonétiques (issues des changements phonétiques réguliers que ces mots ont vécu dans leurs langues respectives). Cependant, les jeux de cognats, des plus classiques comme les listes de Swadesh (1955) à des versions plus récentes comme les initiatives de Dunn (2012) ou Dunn *et al.* (2016), sont de taille trop restreinte (d'une cinquantaine à quelques centaines de paires de mots) pour entraîner des réseaux de neurones. Nous avons donc décidé d'explorer l'utilisation de lexiques bilingues génériques, qui, pour des langues proches, incluent forcément des paires de cognats (en un nombre possiblement plus important que les jeux de référence [3]) au sein d'une majorité de paires qui, ne l'étant pas, constituent du bruit pour notre tâche.

Pour ces expériences préliminaires, nous nous sommes donc intéressés à deux types de données : des

3. Les jeux de références contiennent, au grand maximum, quelques centaines de paires de mots comme cognats attestés. On peut supposer que dans un lexique bilingue d'au moins une dizaine de milliers de mots se trouvent, en plus de ceux-ci, des cognats non attestés pour le moment, mais qui, par leur nature, contiennent les changements phonétiques que l'on cherche à étudier. Ce nombre est malheureusement difficilement quantifiable.

Langues	PL-CZ	PL-LT	PL-IT
Jeu d'entrainement (lexique bilingue)	13,216	6,290	17,158
Jeu de test (paires de cognats)	370	47	57

TABLE 1 – Nombre de paires de mots par jeu de données

jeux de cognats (jeux de test), et des lexiques bilingues (jeux d'entraînement), entre du polonais (PL) et, de la langue la plus proche à la moins proche, du tchèque (CZ), du lituanien (LT), et de l'italien (IT).

3.1.1 Extraction et pré-traitement

EtymDB (Sagot, 2017) est une base de données étymologique extraite automatiquement du Wiktionary, comprenant des lexèmes (triplets de la forme ⟨langue, lemme, sens représenté par une ou plusieurs gloses en anglais⟩) reliés par différentes relations étymologiques typées, dont la relation « hérité de ». Pour générer les paires de cognats pour nos premières expériences, nous suivons les chemins entre les mots, considérant que deux d'entre eux sont cognats si jamais ils partagent un ancêtre dans une de leurs langues parentes communes, et en descendent en ligne droite [4].

YaMTG (Hanoka & Sagot, 2014) est un graphe de traduction multilingue et libre extrait automatiquement du Wiktionary, et une des rares bases de données libres contenant nos langues d'intérêt. Nous en extrayons les entrées bilingues pour nos trois paires de langues, pour créer trois lexiques bilingues (après en avoir retiré les paires de mots contenant des caractères inattendus).

Ces deux jeux sont ensuite phonétisés avec Espeak (Duddington, 2015). Dans le cas où des paires de mots contiennent un lexème identique à l'entrée mais des traductions différentes dans la langue de sortie, ne sont conservées que les paires avec la distance de Levenshtein la plus courte (méthode permettant le meilleur rappel de cognat d'après List *et al.* (2017)).

3.1.2 Propriétés

Les jeux d'entraînement résultants (Table 1) comprennent environ 13 000 paires entre le polonais et le tchèque, 6 000 entre le polonais et le lituanien, et 17 000 entre le polonais et l'italien. Les jeux de cognats sont en moyenne 100 fois plus petits : 370 paires entre le polonais et le tchèque, contre seulement 47 entre le polonais et le lituanien, et 57 entre le polonais et l'italien.

4. Les ancêtres communs présents dans la base sont le slave commun, le proto-balto-slave et le proto-indo-européen pour le polonais et le tchèque, les deux derniers pour le polonais et le lituanien, et seulement le proto-indo-européen pour le polonais et l'italien.

3.2 Paramètres expérimentaux

Après des expériences préliminaires, nous avons déterminé que les meilleurs paramètres pour le modèle neuronal, dans le cadre de cette expérience, étaient l'utilisation d'une couche cachée de type Gated Recurrent Units (GRU) de dimension 100 pour l'encodeur et le décodeur. Les décodeurs utilisent de plus l'attention « dot » de Luong (Luong *et al.*, 2015). L'optimiseur est de type Adam, avec un taux d'apprentissage de 0,001.

Pour le modèle statistique, les paramètres utilisés sont ceux décrits précédemment.

Le jeu d'entraînement est séparé en 70%-30%, 80%-20%, et 90%-10% pour les étapes, respectivement, d'apprentissage et de mise au point. Chaque séparation est aléatoirement générée 10 fois, de façon à observer l'impact de la variation des tailles respectives de ces deux jeux sur l'entraînement. Dans notre cas, cette séparation n'ayant eu aucun impact statistiquement significatif sur les résultats, ceux-ci sont tous traités ensemble ici.

3.3 Résultats

LANGUES	PL→CZ	PL→LT	PL→IT
	Résultats sur les jeux de tests		
MEDeA	48.44 ± 1.13	19.22 ± 0.99	26.75 ± 1.05
Moses	54.43 ± 0.41	20.55 ± 0.62	28.37 ± 0.62

TABLE 2 – Scores BLEU de MEDeA et Moses (moyenne sur 30 expériences).

En observant la Table 2, le premier constat que nous faisons est que tous nos résultats sont masqués par une contrainte matérielle : la quantité de données d'entraînement. En effet, le jeu PL-LT est deux à trois fois plus petit que ses confrères, et il est le jeu avec les plus mauvais résultats. Cependant, si seule la taille des données avait un impact sur l'apprentissage, on s'attendrait à ce que le jeu le mieux prédit soit celui entre l'italien et le polonais, soit le plus gros jeu, alors que les meilleurs résultats sont ceux obtenus entre le polonais et le tchèque, de 20 points BLEU meilleurs pour un jeu d'entraînement 30% plus petit. Nous supposons donc que, sous réserve d'avoir suffisamment de données, plus des langues sont proches, plus leurs correspondances sont faciles à apprendre.

Le second constat est que la méthode neuronale est systématiquement moins bonne que la méthode statistique sur la tâche de prédiction de cognats après un entraînement sur des lexiques bilingues. Ceci peut indiquer soit que les méthodes neuronales de type encodeur-décodeur ne sont pas adaptées à cette tâche dans l'absolu, soit que les méthodes statistiques ont une meilleure capacité à extraire des informations de données très bruitées. Nous notons également que les résultats ne sont en moyenne pas très bons, et que la piste des lexiques bilingues ne semble pas être aussi intéressante qu'elle aurait pu. Pour différencier entre ces hypothèses, nous proposons les expériences suivantes.

4 Les méthodes neuronales conviennent-elles à l'apprentissage de correspondances phonétiques ? Expériences sur des données artificielles

4.1 Contexte

L'utilisation de lexiques bilingues pour apprendre à prédire des cognats s'est avérée être une piste peu satisfaisante. Cependant, deux raisons majeures peuvent être à l'origine des difficultés rencontrées par nos modèles : soit les réseaux de neurones ne sont pas aussi bien adaptés que les méthodes statistiques à l'apprentissage de ces changements, soit ils sont plus sensibles au bruit que des méthodes statistiques, et les lexiques bilingues constituaient des données trop bruitées pour l'apprentissage de notre tâche.

Il est possible d'invalider facilement une des hypothèses évoquées précédemment : si les réseaux de neurones ne conviennent pas à l'apprentissage des changements phonétiques, alors ils n'apprendront jamais aussi bien ou mieux que des méthodes statistiques, même sur des données parfaites. Par contre, s'ils sont capables d'apprendre sur des données parfaites, alors le problème vient plus probablement de leur sensibilité au bruit que de leur inadéquation à la tâche.

Pour comprendre ce qu'il est effectivement possible d'apprendre ou non, nous décidons de générer des lexiques phonétiques artificiels, composé d'une proto-langue, et de deux langues filles générées en appliquant à la langue mère des changements phonétiques réguliers. Cette méthode présente deux avantages : elle permet d'étudier la taille minimale de données nécessaires pour apprendre, et de maîtriser complètement les paramètres liés aux données elles-mêmes, de la richesse à la quantité de bruit. Cependant, il est important que les données, bien qu'artificielles, obéissent à des règles de construction et d'évolution réalistes, pour que les résultats des expériences puissent être transposables à des données réelles.

4.2 Données artificielles

Nous choisissons pour cela de créer une proto-langue à partir d'un inventaire de phones et d'une phonotactique (organisation syllabique des sons dans la langue), puis d'en dériver les langues filles à partir de l'application séquentielle de changements phonétiques plausibles.

Nous avons ainsi développé un algorithme, qui, à partir d'un inventaire de phones et d'une phonotactique, génère un lexique d'une taille choisie. Dans le cadre de ces expériences, nous choisissons de nous inspirer du latin pour la proto-langue, et des langues romanes pour ses langues filles. Nous utilisons donc trois sources historiques. Pour la génération de la proto-langue (PL) nous utilisons tout d'abord l'inventaire phonétique des langues romanes : chaque lexique généré en utilise les phones communs à toutes les langues et tire aléatoirement un sous-ensemble des phones moins courants. Nous utilisons ensuite une version simplifiée de la phonotactique du latin classique (inspirée de (Cser, 2016)), qui nous permet de générer des mots à partir d'un nombre de syllabes aléatoirement choisi, lesquelles sont construites en suivant les règles phonotactiques liées à leur emplacement au sein des mots. Enfin, la dernière source sont les changements phonétiques des langues romanes. Pour générer les langues filles, l'algorithme choisit aléatoirement un sous-ensemble de changements phonétiques parmi ceux possibles (dont l'apocope, l'épenthèse, la palatalisation, la lénition, la prothèse de voyelles et la diphtongaison), puis les applique successivement au lexique de la proto-langue pour générer le

lexique d'une langue fille.

Pour nos expériences, nous avons finalement généré, à partir d'une proto-langue (PL), deux langues filles (F1 et F2) avec 15 changements phonétiques chacune, soit 20 000 triplets de mots. Voici deux exemple du type de données phonétisées obtenues : $[\text{stra}]_{\text{PL}} > [\text{isdre}]_{\text{F1}}$, $[\text{estre}]_{\text{F2}}$ et $[\text{ʒolpast}]_{\text{PL}} > [\text{ʒolbes}]_{\text{F1}}$, $[\text{ʒolpes}]_{\text{F2}}$

4.3 Paramètres expérimentaux

Notre but lors de ces expériences est double : déjà, vérifier si les méthodes neuronales peuvent apprendre, mais également, le cas échéant, déterminer si la quantité de données a un impact sur les performances relatives des deux types de modèles. Pour cette raison, nous réitérons les expériences pour différentes tailles de jeux de données : 500, 1000, 1500, 2000 et 3000 triplets de mots, divisées en 80% pour l'entraînement et 20% pour le test. Ces données sont choisies aléatoirement parmi les 20 000 triplets générés précédemment, selon 3 graines d'aléa (« *random seeds* ») différentes.

En ce qui concerne les modèles statistiques, 80% du jeu d'entraînement est utilisé pour leurs étapes d'apprentissage et 20% pour leurs étapes de mise au point ; un modèle statistique différent doit être entraîné par paire de langue possible.

Le modèle neuronal, quand à lui, utilise toutes les données d'entraînement pour l'apprentissage, de toutes les langues à toutes les langues en une seule fois ; la taille de couche cachée donnant les meilleurs résultat était de 25 après des expériences préliminaires [5].

Les deux modèles prédisent de la meilleure aux trois meilleures réponses.

4.4 Résultats

Sur les données synthétiques dénuées de bruit, le modèle statistique comme le modèle neuronal apprennent très bien à prédire d'une langue fille (F1) à une autre langue fille (F2), et atteignent des scores très nettement supérieurs à ceux obtenus lors des expériences préliminaires sur les lexiques bilingues : entre 88 et 97 BLEU pour MEDeA et 90 à 96 pour Moses, soit des résultats équivalents, comme on peut le voir sur les colonnes F1-F2 et F2-F1 de la figure 1. Les réseaux de neurones peuvent donc apprendre à prédire des cognats d'une langue fille à une autre langue fille, à partir de données de cognats, si tant est que ces données soient de qualité et de quantité suffisante.

4.4.1 Impact de la direction de prédiction sur les résultats

Cependant, nous notons également que toutes les directions de prédictions ne sont pas équivalentes. Prédire de la proto-langue (PL) à une langue fille (F) donne les meilleurs résultats (de 94 à 99 BLEU), tandis que prédire d'une langue fille à la langue mère est, de très loin, la tâche la plus difficile (de 60 à 80 BLEU). De plus, prédire la deuxième et la troisième meilleure réponse augmente considérablement les scores BLEU dans le cas des situations présentant une ambiguïté forte (F→PL), et ce quel que soit le modèle considéré.

5. La différence considérable de taille de la couche cachée entre cette expérience et la précédente s'explique par la différence de taille entre les données utilisées, d'un facteur de 3 à 20.

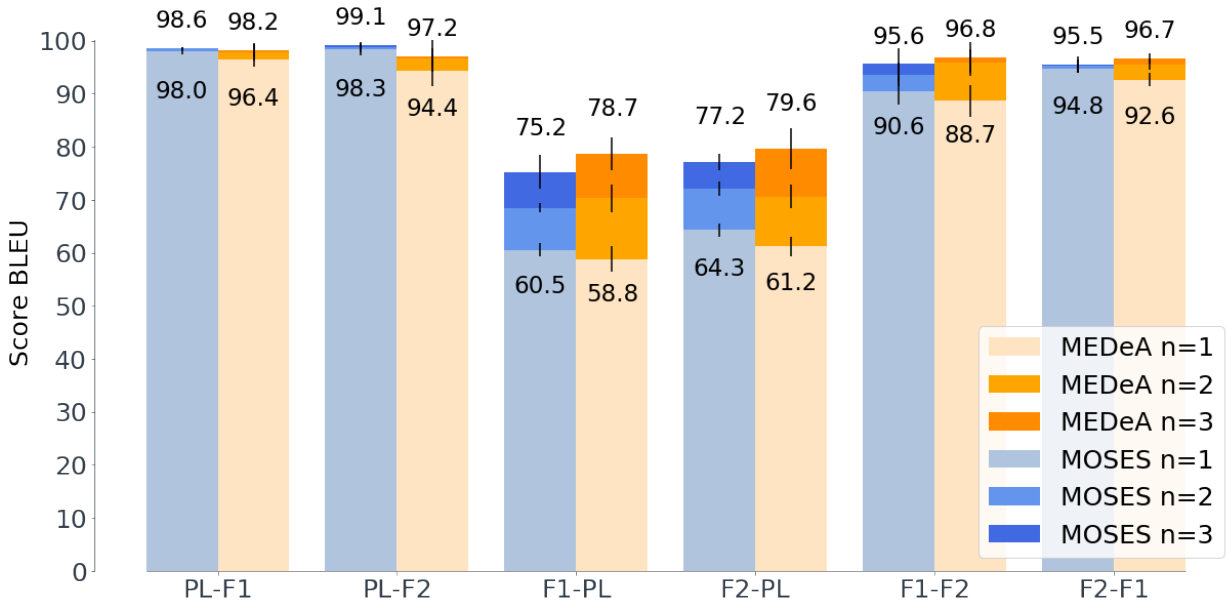
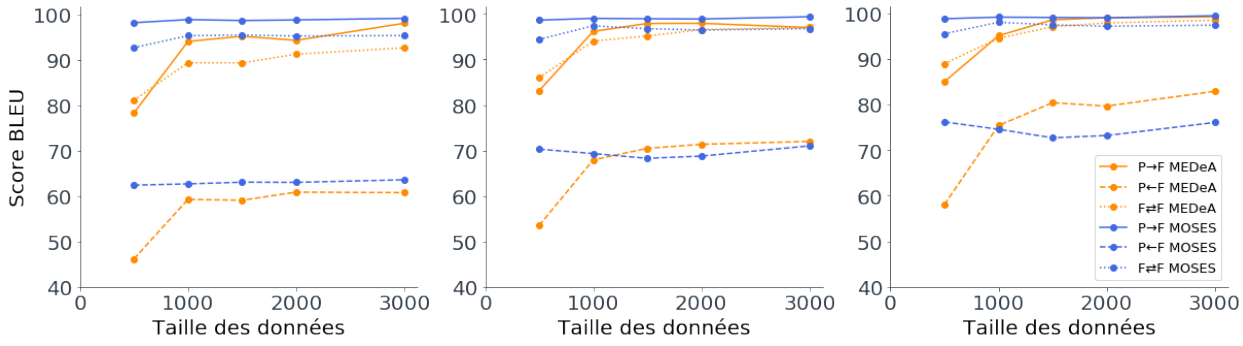


FIGURE 1 – Scores BLEU pour les n meilleures prédictions, à partir de 1000 paires de mots (en fonction de la direction de prédiction).

4.4.2 Impact de la taille des données et du nombre de prédictions sur les résultats



(a) n=1 (b) n=2 (c) n=3

FIGURE 2 – Scores BLEU pour les n meilleures prédictions.

Le modèle statistique est systématiquement meilleur que le modèle neuronal quand on calcule le score BLEU sur la meilleure réponse seulement, encore que les performances se rejoignent quand on dépasse 2000 triplets de mots pendant l'entraînement, comme on peut le voir dans la figure 2.

Par contre, pour les deux ou trois meilleures réponses, les résultats changent : le modèle statistique est significativement meilleur pour de très petites tailles de données (500 triplets) [6], mais le modèle neuronal est bien meilleur pour un grand nombre de données (2000 triplets et plus). Sur la zone intermédiaire, entre 1000 et 1500 triplets, les performances sont à peu près équivalentes entre les modèles.

Cette expérience nous a donc permis de montrer que, pour des données idéales, un modèle neuronal

6. On suppose que le réseau de neurone est moins performant pour de petites tailles de données car il sur-apprend.

est aussi performant qu'un modèle statistique à partir de 1500 paires de mots par couple de langue, notamment quand on s'intéresse aux situations impliquant de l'ambiguïté. Qu'en est-il sur des données réelles ?

5 Peut-on apprendre des changements phonétiques sur des données réelles ?

Pour déterminer si nos résultats sont généralisables, nous menons les mêmes expériences sur des données réelles. Il est attendu que ces expériences fonctionneront moins bien que les précédentes, dans la mesure où les données réelles contiennent du bruit, complètement absent de nos données artificielles.

5.1 Données réelles

Nous avons besoin de travailler sur un jeu de cognats de langues réelles liées impliquant une langue parente bien connue et plusieurs langues filles proches mais différentes. Nous choisissons d'étudier le latin (LA) comme langue mère et l'italien (IT) et l'espagnol (ES) comme langues filles.

Pour extraire nos jeux de données, nous utilisons EtymDB 2.0 (Fourrier & Sagot, 2020), la version la plus récente de la base de données EtymDB présentée à la section 3.1.1. Nous utilisons le même raisonnement pour extraire les cognats : 2 mots sont cognats s'ils partagent un ancêtre pour une de leurs langues parentes communes (ici, le latin, le latin pré-classique, le proto-italique et le proto-indo-européen pour les paires LA-IT et LA-ES, et ces langues plus le latin vulgaire pour la paire IT-ES). La phonétisation et le pré-traitement sont de nouveau effectués avec Espeak puis l'application d'une distance de Levensthein pour retirer les doublons erronés. Toute séquence de phone est précédée d'un token de début de phrase indiquant sa langue, et terminée par un token de fin de phrase.

Le jeu final contient 605 triplets LA-ES-IT, qui ont été manuellement ré-éxaminés, et serviront principalement de jeu de test (2/3 de test, 1/3 de train). Les données d'entraînement sont constituées de 5040 paires de cognats pour IT-LA, 4208 pour ES-LA, et 1801 pour ES-IT, desquelles sont retirées les données de test.

5.2 Paramètres expérimentaux

MEDeA est entraîné avec toutes les données sur toutes les combinaisons possibles entre les langues (IT, ES, LA), pendant 50 itérations (« *epochs* »). Nous l'entraînons sur 3 graines d'aléa, et comparons des tailles de couche cachée entre 12 et 50, et la précision de la meilleure aux trois meilleures prédictions. Après observations, nous constatons que pour ces jeux, la meilleure taille de couche cachée est de 37, avec des plongements vectoriels de taille 10, et ce sont les valeurs que nous utiliserons pour la suite des expériences.

Moses est entraîné sur les différentes combinaisons de paires de langues séparément, avec les mêmes divisions dans les données. Les données triples sont traitées comme des combinaisons de paires.

5.3 Résultats préliminaires

Nous observons, durant ces expériences, que le modèle statistique est systématiquement meilleur que le modèle neuronal, d'environ 15 points. Les réseaux de neurones sont très sensibles au bruit et aux incohérences dans les données.

Cependant, là où le modèle statistique obtient pour l'instant une performance absolue supérieure, le modèle neuronal semble apprendre une structure sous-jacente des données, qui lui permet de mieux gérer les cas d'ambiguïté. En effet, pour chacune des paires de mots de nos jeux de test, nous associons à une entrée unique une sortie unique. Ainsi, quand un modèle prédit plusieurs réponses, une seule sera correcte *par rapport au jeu de données*. Cependant, les autres ne seront pas pour autant fausses dans l'absolu, et on peut distinguer 3 cas :

1. le modèle prédit des réponses secondaires historiquement valides (par exemple, le même adjectif, accordé différemment)
2. le modèle prédit des réponses secondaires historiquement plausibles (par exemple, un nom dans un autre genre, incorrect mais plausible par rapport à la structure de la langue)
3. le modèle prédit des réponses secondaires complètement incorrectes

Observons quelques exemples de l'italien au latin. Le réseau de neurone prédit à plusieurs reprises des formes grammaticalement valides, comme [rustiko] 'rustique', venant de [rʊstɪkʊs] 'de la campagne', qui voit son ancêtre prédit comme étant [rʊstɪkʊs] (masc. - bonne réponse), [rʊstɪkʊm] (neut. — cas 1), ou [rʊstɪkss] (aucun sens — cas 3) par MEDeA, contre [rʊkʊstɪ], [rʊɪkɔst] ou [ʊsrtɪkwʊs], trois formes dépouvues de sens, par Moses (toutes cas 3). MEDeA nous a d'ailleurs permis d'identifier des erreurs dans nos jeux de données : [ramo] 'branche' $<$ [ramʊs] 'branche', était relié de façon erronée à [radɪks] 'racine' (qui est un cognat de [ramʊs]) ; MEDeA a prédit cet ancêtre comme étant [ramʊs] (masc. — bonne réponse), [ramʊ] (cas 3), ou [ramʊm] (forme neutre du nom, incorrecte, mais plausible, soit le cas 2), tandis que Moses a prédit [mʊr], [rɛam], ou [raɛm] (toutes le cas 3 à nouveau).

Le modèle statistique produit donc plus souvent *la* bonne réponse par rapport à nos données, obtenant ainsi un meilleur score, là où le modèle neuronal produit plus souvent plusieurs réponses plausibles.

6 Conclusion

Dans cet article, nous avons d'abord montré que les lexiques bilingues, bien que semblant au premier abord convenir à la tâche d'apprentissage de changements phonétiques (car porteurs de cette information et de taille raisonnable), étaient en réalité trop bruités, pour les réseaux de neurones comme les méthodes statistiques étudiées. Nous avons ensuite montré, en travaillant sur des données artificielles, que ces deux types d'algorithmes présentent des forces et faiblesses complémentaires sur des jeux de taille réaliste sans être trop restreints ; les méthodes statistiques sont meilleures dans les cas ne présentant pas d'ambiguïté (direction de prédiction chronologique), et les réseaux de neurones dans les cas en présentant beaucoup (direction de prédiction chronologique inverse). Enfin, en cherchant à confirmer ces résultats sur des données réelles, nous avons montré que les méthodes statistiques sont, avec les paramètres choisis pour ces expériences, plus performantes que les méthodes neuronales, mais que ces dernières semblent faire de meilleures généralisations sur les données. Des expériences complémentaires restent à faire. Une première étape serait la création des données artificielles bruitées, ou subissant des changements phonétiques plus complexes, pour

étudier la performance respective des deux types de modèles sur celles-ci. Une seconde est l'étude plus détaillée des apprentissages des différents modèles, à l'échelle du mot et du phone ; pour ce faire, il pourrait être intéressant de chercher une métrique plus pertinente pour la tâche de linguistique historique que le score BLEU, qui pénaliserait moins les phones prédits quand ils sont proches de ceux attendus [7].

Références

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

BEINBORN L., ZESCH T. & GUREVYCH I. (2013). Cognate production using character-based machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, p. 883–891, Nagoya, Japan : Asian Federation of Natural Language Processing.

CSER A. (2016). *Aspects of the phonology and morphology of Classical Latin*. Thèse de doctorat, Pázmány Péter Katolikus Egyetem. DOI : 10.1111/1467-968X.12184.

DEKKER P. (2018). Reconstructing language ancestry by performing word prediction with neural networks. *Master. Amsterdam : University of Amsterdam*. DOI : 10.13140/RG.2.2.32990.33601.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DUDDINGTON J. (2015). eSpeak : Text to speech. http://espeak.sourceforge.net/index.html.

DUNN M. (2012). IELex : Indo-European Lexical cognacy database. http://ielex.mpi.nl/.

DUNN M., GARGETT A., RUNGE J. & KHAIT I. (2016). CoBL : Cognacy in Basic Lexicon. https://github.com/lingdb/CoBL-public.

FOURRIER C. & SAGOT B. (2020). Methodological Aspects of Developing and Managing an Etymological Lexical Resource : Introducing EtymDB-2.0. In *Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France.

HANOKA V. & SAGOT B. (2014). YaMTG : An Open-Source Heavily Multilingual Translation Graph Extracted from Wiktionaries and Parallel Corpora. In *Language Resources and Evaluation Conference*, Reykjavik, Iceland : European Language Resources Association. HAL : hal-01022306.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R. *et al.* (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, p. 177–180. DOI : 10.5555/1557769.1557821.

LIST J.-M., GREENHILL S. J. & GRAY R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLOS ONE*, **12**(1), 1–18. DOI : 10.1371/journal.pone.0170046.

LUONG T., PHAM H. & MANNING C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1412–1421, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1166.

7. Prédire une autre voyelle à la place d'un [a] mérite d'être moins pénalisé que de prédire une consonne, par exemple.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51. DOI : 10.1162/089120103321337421.

OSTHOFF H., OSTHOFF H. & BRUGMANN K. (2014). In *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen (reprinted)*, volume 4 de *Cambridge Library Collection - Linguistics*, p. 418–418. Cambridge University Press. DOI : 10.1017/CBO9781139600132.006.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphie, Pennsylvanie, USA. DOI : 10.3115/1073083.1073135.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6319.

SAGOT B. (2017). Extracting an Etymological Database from Wiktionary. In *Electronic Lexicography in the 21st century (eLex 2017)*, p. 716–728. HAL : hal-01592061.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*. DOI : 10.5555/2969033.2969173.

SWADESH M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, **21**(2), 121–137. DOI : 10.1086/464321.