Slide d'ouverture:

Bonjour, je vous remercie de votre invitation. Je commence par préciser qu'une partie des éléments techniques plus complexes de cette communication ne sont pas de moi, vous trouverez les liens vers les outils que j'utilise en fin de présentation.

Cette présentation vise à expliquer comment nous pouvons essayer d'évaluer l'évolution de la façon de traduire Homère en France. Cela implique que nous soyons capable d'évaluer automatiquement le sens des mots en contexte pour le grec et le français. Le sens des mots changeant avec le temps, nous devons essayer de déterminer diachroniquement le sens de tel ou tel terme chez tel ou tel traducteur pour traduire tel ou tel mot en grec. Je m'intéresse ici uniquement à l'*Odyssée*.

Slide 1 :

Pour ma thèse, soutenue en 2018, j'ai tenté de comparer une cinquantaine de traductions en langue française (mais pas que, j'avais aussi du latin, de l'anglais et de l'italien), depuis celle de Peletier du Mans jusqu'à celle de M. Mugler en 1991. Plusieurs écueils se présentaient donc d'emblée.

Slide 2 :

Premier écueil bien sûr, le nombre de traductions à prendre en compte. Rapidement, un panel non exhaustif de celles qui sont prises en compte dans l'expérience que je vais vous montrer tout à l'heure.

Les deux premiers chants de Peletier du Mans, 1540 ; Amadis Jamyn, 1584 ; Salomon Certon, 1604 ; André Boitel, 1619 ; Achille de La Valterie, 1681 ; Anne Dacier, 1716 ; Rochefort, 1777 ; Bitaubé, 1785 ; Lebrun en 1819 ; Dugas-Montbel en 1833 ; Bignan, 1841 ; Eugène Bareste en 1842 ; Paul Giguet en 1852 ; Emile Pessonneaux en 1862 ; Leconte de Lisle en 1867 ; Froment en 1883 ; Hins en 1883 aussi ; Sommer en 1886 (et plus encore) ; Séguier en 1894 ; Calbet Rosny en 1897 ; Bérard en 1924 ; Mario Meunier en 1943 ; Dufour et Raison en 1946 ; Jaccottet en 1955 et F. Mugler en 1991.

Autre écueil évident : le texte source : même si le texte grec de l'*Odyssée* est « relativement » stable, il faut pouvoir tenir compte des variantes (c'est là que le format numérique est d'un grand secours, puisque l'on peut prendre en compte, en un seul et même document, l'ensemble des leçons).

Slide 3 :

Voilà comment j'avais procédé durant la thèse (à savoir que j'emploie maintenant des méthodes différentes, plus poussées sur lesquelles je reviendrai). La première tâche de l'ordinateur pour comparer les traductions est proche de celle du traducteur : il faut pouvoir déterminer quelle portion de sens dans la source correspond à telle ou telle portion de la cible. Les textes sacrés par exemple sont plus simples à aligner parce que naturellement découpés et alignés. Pour l'*Odyssée*, déterminer des unités de sens et en trouver la traduction peut être très compliqué : la source peut être traduite en plusieurs séquences françaises, et vice versa.

Pour trouver une solution à ce problème, j'ai employé des algorithmes empruntés à la bio-informatique. Si l'on se figure que chaque élément de sens de la source équivaut à une information génétique, il est possible, comme pour l'ADN, de combler les séquences manquantes pour associer les chaînons.

Slide 4:

Comment déterminer les séquences de sens ? J'ai (dans un premier temps, mais je n'en ai plus besoin maintenant) repéré des pivots stables, présents majoritairement à la fois dans la source et dans la cible, à savoir les noms propres. J'ai donc découpé les textes en séquences, en déterminant qu'une séquence nouvelle commençait à chaque fois que l'algorithme rencontrait un nouveau nom. Cela suppose un prétraitement de chaque texte, avec lemmatisation et étiquetage de nature (si vous avez besoin d'outils sur ce point, je peux vous orienter).

Cela fait, il va falloir tenter de déterminer le degré de similarité de sens entre une séquence source et une séquence cible. Pour la thèse j'ai cumulé une série de mesures, que vous pouvez voir à l'écran, comme la proximité syntaxique de certains éléments, des dictionnaires de distribution etc.

Slide 5 :

Une fois que nous avons obtenu un degré de similarité entre chaque séquence source face à chaque séquence cible, il faut choisir le chemin optimal qui permet de reconstituer l'alignement de l'intégralité de la cible en fonction de la source. C'est là que

la bio-informatique intervient, avec notamment l'algorithme de Needleman Wunsch.

Slide 6 :

Voici ce que j'ai obtenu (il reste des erreurs et l'alignement n'est pas parfait) dans un premier temps. Rapidement je vous explique ce que montre la démonstration en ligne (encore une fois désolée pour les coquilles restantes). Explication.

Slide 7 :

Un exemple de ce que j'ai pu obtenir grâce à un tel outil : le cas du texte à plusieurs mains avec Dufour et Raison.
Par courtoisie et en hommage à son défunt collègue, Jeanne Raison indique dans son propos liminaire que la traduction aurait dû être signée d'un seul nom. Elle laisse à penser que l'ensemble de la traduction, à l'exception de certains passages, avait déjà été constituée par Dufour. Notre hypothèse est que Dufour a traduit l'Odyssée jusqu'au chant XVIII inclus, et que Raison a dû prendre le relais à partir du chant XIX. Nous avons été amenée à formuler cette hypothèse en comparant les sources des différents chants, et nous avons trouvé un changement important dans les sources employées par les traducteurs. En effet, avant le chant XVIII, Sommer n'est jamais présent dans les sources de la traduction. Dufour ne semble pas s'appuyer sur Sommer (ce qui pourrait, toutes précautions gardées, ne pas surprendre : Sommer est un traducteur scolaire à usage des petites classes, Dufour est un professeur d'université, plus proche de Bérard, tant du point de vue du statut que de l'approche philologique). En revanche, à partir du chant XIX, Bérard n'apparaît plus du tout comme source, à l'exception des chants XXII et XXIII, et semble massivement remplacé par Sommer, qui gagne presque toujours la deuxième place dans le pourcentage de reprises de basses fréquences, après Meunier. Ce premier indice nous a poussée à vérifier dans les détails la correspondance des systématismes homériques avant et après le chant XIX. Notre hypothèse est que Dufour a traduit intégralement l'Odyssée jusqu'au chant XIX, qu'il n'a pas traduit les chants XIX à XXI, qu'il a partiellement traduit les chants XXII et XXIII, et qu'il n'a pas traduit dans son entier le chant XXIV. Effectivement, si les noms sont scrupuleusement rendus de la même manière avant et après le chant XIX, il n'en est rien pour

les épithètes, qui sont systématiquement les mêmes avant le chant XIX, et radicalement différentes à partir du chant XIX. Par exemple, pour l'expression « ὑλήεις Ζάκυνθος », nous trouvons systématiquement avant le chant XIX la traduction « Zacynthe couverte de forêts », tandis qu'après le chant XIX l'expression est traduite par « Zacynthe boisée ». De même pour l'expression « περίφρων Πηνελόπεια », qui apparaît à six reprises des chants I à XVIII sous la forme de « sage Pénélope », à plus de vingt reprises avant le chant XIX, et qui disparaît dans les chants XIX à XXI ainsi que dans le chant XXIV, pour être remplacée par l'expression systématique dans ces chants de « prudente Pénélope ». Il y a d'autres exemples, mais je ne développe pas. Les exemples de ce type sont nombreux et systématiques. L'analyse des sources potentielles de chaque traducteur prouve que notre outil informatique peut servir aussi à l'attribution auctoriale de traductions à plusieurs mains.

Slide 8 :

Mais il est désormais possible d'aller plus loin, et d'utiliser des outils qui n'existaient pas il y a peu. La méthode que nous utilisions ici était relativement peu adaptable : les mesures de similarité ont bien fonctionné pour notre objet probablement moins pour d'autres textes, et plus encore le découpage en séquences nécessite une stabilité de la structure textuelle très contraignante. L'IA actuelle permet de s'affranchir de ces liens, et il est possible d'espérer identifier des proximités sémantiques évolutives entre le grec et le français. Théoriquement, cette approche nous permettrait de voir non seulement comment tel ou tel terme est traduit à un temps T, mais aussi comment ce terme évolue parmi les autres, par quels termes il est attiré, ou repoussé.
Comment fait-on pour identifier le sens d'un mot avec un ordinateur ?

Slide 9 :

Il faut essayer d'imaginer que le langage est un espace géométrique. Théoriquement il a autant de dimensions qu'il y a de mots, mais ici contentons nous d'un espace en deux dimensions. Chaque mot aurait une place dans l'espace en fonction des mots qui l'entourent. Ainsi, des mots qui apparaîtraient dans le même contexte seraient proches dans l'espace sémantique. On représente

alors mathématiquement des mots comme des vecteurs dans l'espace sémantique. Ici par exemple, nous voyons que sur un espace en deux dimensions, théoriquement, gunè et anèr sont plus proches dans l'espace que turannos et basileia.
Pour évaluer la proximité des vecteurs, on mesure les angles entre eux (ce qu'on appelle la similarité cosinus). Dans l'exemple, c'est ce degré de similarité, c'est à dire de proximité d'emploi par rapport à l'ensemble du corpus, qui nous permet de dire que gunè et anèr sont proches dans l'espace sémantique.
Ce type d'approche permet aussi de déduire des identités de rapport entre les termes. Par exemple, l'angle entre gunè et anèr peut nous permettre de déduire basileia. Comment ? Il suffit d'appliquer l'angle entre anèr et gunè au vecteur turannos pour obtenir basileia. En d'autres termes, nous pouvons déduire que anèr est à gunè ce que turannos est à basileia (mathématiquement en soustrayant et ajoutant les vecteurs les uns aux autres, je développerai si besoin).

Slide 10 :
Ici c'est la représentation d'un espace vectoriel monolingue en trois dimensions (entraîné sur trois œuvres de Dumas) : chez Dumas, nous pouvons voir quels sont les termes proches de « roi ». Les résultats obtenus sont bien sûr hautement dépendants du corpus d'entraînement.

Slide 11 :
Mais cela ne résout pas un problème essentiel pour nous : comment faire un espace vectoriel en plusieurs langues, puisque typiquement les mots ne peuvent pas d'emblée, être proches par leur contexte ?
Depuis très peu de temps (2017), la majorité des traducteurs automatiques utilisent un système dit neuronal. La traduction neuronale fonctionne comme sur cette illustration (très simplifiée). Par rapport aux techniques plus anciennes, et grâce aux progrès dans la puissance de calcul, la machine neuronale permet d'engranger beaucoup plus d'informations sur chacun de mots qu'elle doit traduire. Le principe est simple : elle transforme chaque mot en vecteur de *n* dimensions, l'inclut dans un espace vectoriel, et grâce à des données parallèles de masse, adapte cet espace source à un même espace vectoriel cible. Problème de ce

type d’approche pour nous : la machine neuronale a besoin d’énormément de données parallèles, c’est-à-dire de corpus découpés en micro-séquences dont on sait que telle séquence source correspond à telle séquence cible. Clairement ce sont des données qui n’existent pas pour les langues anciennes.
On a encore une chance, celle des traductions existantes. Même si nos traductions sont approximatives, même si elle ne sont pas alignées séquence à séquence, même si elles sont lacunaires, elle peuvent nous servir. Pourquoi ? Parce qu'on sait deux choses : d'une part qu'une traduction parle peu ou prou de la même chose que le texte qu'elle traduit ; et d'autre part parce qu'un traducteur a fait l'effort de créer un pont entre les cultures et entre les langues, et que donc ces notions charnières sont, même imparfaitement, transposables, même potentiellement dans des contextes différents.
Mais même aligner des traductions est un problème très compliqué. Pourquoi ? Parce que déjà justement les traducteurs peuvent ne pas avoir la même approche de leur art : certains vont coller à leur source, d'autres non, etc. Et d'autre part, parce que si l'on veut aller jusqu'à aligner des mots, nous allons rencontrer des problèmes structurels de la langue, comme le fait qu'à un mot source corresponde plusieurs mots cible etc.
Comment fait-on ? Comment faire pour qu'une machine apprenne les équivalences sémantiques sans qu'on lui ait au minimum donné des équivalences de base ? Il y a un principe fondateur qui semble aller de soi, mais qui en réalité est déjà un présupposé important, même dans le cas des traductions : les langues fonctionnent sémantiquement de la même manière, et même distributionnellement de la même manière. Autrement dit, géométriquement parlant, les langues sont isomorphes. Donc, quand on a un texte traduit et sa source, on suppose donc non seulement que le contenu est le même, mais que quelque part la façon de transmettre ce contenu est la même aussi.
L’idée est donc de créer deux espaces sémantiques bien séparés, et d’essayer de les faire coïncider.
Je vous fais grâce des explications précises mathématiques sur la manière dont on va superposer les espaces : si vous le souhaitez je pourrai vous expliquer.

Le principe est de faire une rotation de la source pour qu'elle se superpose un minimum à la cible. Pour ce faire,à ma connaissance, deux techniques existent.

Dans la première on utilise ce qu'on appelle l'"adversarial training" : l'adversaire essaie de tromper le discriminateur qui, apprenant de ses erreurs, s'améliore. L'adversaire essaie de maximiser l'erreur en transformant les données, avec certaines conditions sur la manière dont il les transforme (par exemple, l'adversaire peut faire des rotations dans le plan du nuage que le discriminateur utilise pour prédire, il conserve ainsi les distances mais pas les directions). Au final le discriminateur est de plus en plus robuste, et ajuste la rotation. Dans la seconde, celle que j'utilise, plus efficace sur des petits espaces, on utilise la transformation orthogonale linéaire (basiquement expliquée ci-dessous).

Donc théoriquement, à l'issue d'un tel entraînement, on se retrouve avec un espace sémantique multilingue à peu près fiable, même avec peu de données.

Slide 16 :

Voici la représentation d'un espace multilingue entraîné sur l'*Odyssée* (avec variantes) et les traductions présentées plus tôt. Nous pouvons voir dans cet exemple quels sont les plus proches voisins sémantiques sur ce corpus global de cœur, et nous retrouvons en première position frèn et hètor (aussi kardia, kèr arrive bien après). Nous pouvons, au cours de l'entraînement, savoir quels sont les contextes qui ont eu le plus de poids pour déterminer cette proximité.

Slide 17 :

Cela fonctionne aussi dans l'autre sens, sur n'importe quel terme : sans aucun entraînement manuel ou pré-traduction, la machine neuronale a pu déterminer que kéras était proche de corne dans l'espace.

Slide 18 :

Nous pouvons aussi voir dans l'espace les particularités de certains auteurs dans l'espace. Par exemple, ici, j'ai cherché les particularités sémantiques de Bérard, et sa répartition dans l'espace, et je suis tombée, entre autres, sur « croiseur », qu'il utilise de façon très similaire à celle de Jaccottet, contrairement à l'ensemble des autres traducteurs. Ce type d'analyse, menée globalement, peut par exemple permettre de déterminer quels sont les traducteurs qui se détachent majoritairement et pourquoi.

Slide 19 :

Encore une fois, il ne s'agit pas d'une étude aboutie, loin de là. Mais qu'apporte cette approche à l'étude traductologique et diachronique que nous espérons mener ? Il me semble que cette nouvelle approche permet plusieurs choses : concrètement cela permet de voir l'évolution des termes dans le temps (des termes qui s'attirent, qui viennent se greffer au grec ou s'en éloigner) etc. Mais pas seulement. Je rebondis sur la question de l'intertextualité dont on parlait hier : si nous élargissons le corpus grec, nous pouvons voir à la fois l'intertextualité « orthographique » mais aussi allusive ou conceptuelle (on peut voir l'évolution, par exemple, du « concept » Achille dans le temps, et ses traductions).

Je vais prendre un exemple concret par rapport à ce que j'avais pu trouver durant ma thèse.

Je propose une distinction entre le plagiat, ou la reprise, et l'inspiration, terme autrement plus difficile à définir.