Babel 2.0.
Où va la traduction automatique ?

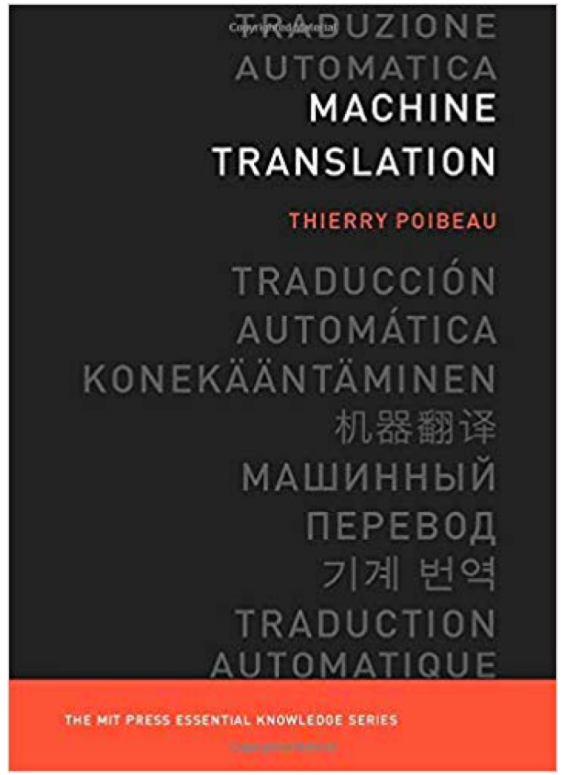
Thierry Poibeau
Laboratoire LATTICE (CNRS & ENS/PSL & Université Sorbonne nouvelle)
PRAIRIE (*Paris Artificial Intelligence Research Institute*)

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

ENS | PSL

cnrs

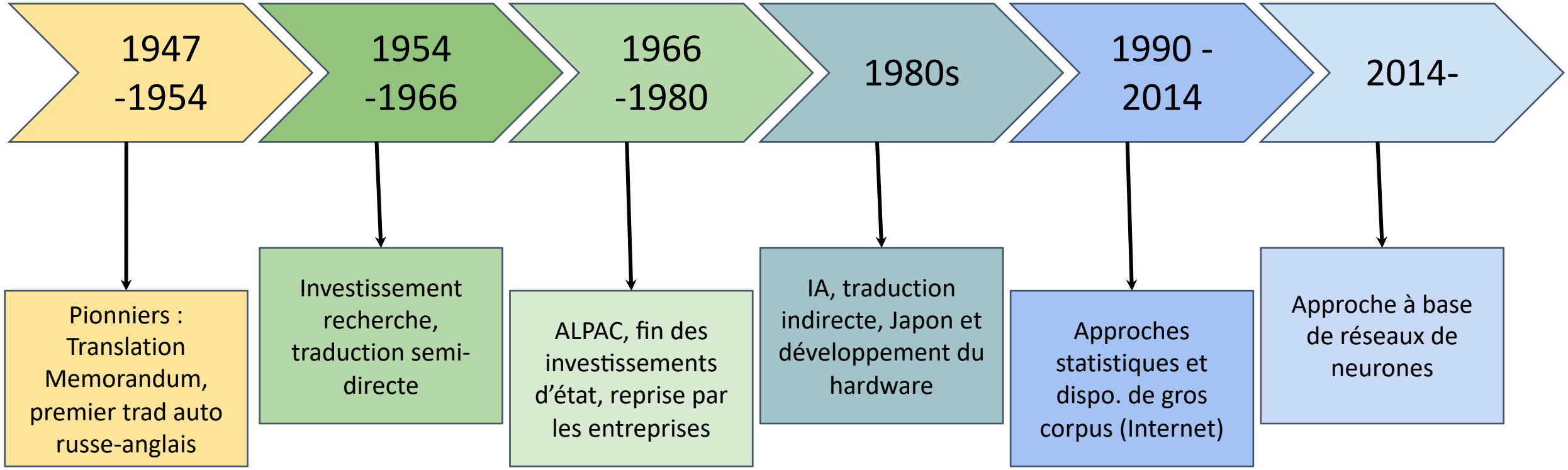

Thierry Poibeau
Babel 2.0
Où va la traduction automatique ?
translate
Odile Jacob


TRADUZIONE AUTOMATICA
MACHINE TRANSLATION
THIERRY POIBEAU
TRADUCCIÓN AUTOMÁTICA
KONEKÄÄNTÄMINEN
机器翻译
МАШИННЫЙ ПЕРЕВОД
기계 번역
TRADUCTION AUTOMATIQUE
THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES

- Quelle est l’histoire de la traduction automatique ? Comment fonctionne la TA aujourd’hui ?
- La TA est-elle aujourd’hui utilisable ? Comment ? Pour quoi faire ? Dans quels contextes ?
- Conclusion

- Quelle est l’histoire de la traduction automatique ? Comment fonctionne la TA aujourd’hui ?
- La TA est-elle aujourd’hui utilisable ? Comment ? Pour quoi faire ? Dans quels contextes ?
- Conclusion

Une brève histoire de la traduction automatique



Adapté d'une présentation de Marianne Reboul (ENS Lyon)

L’intuition de Weaver

- *On peut naturellement se demander si le problème de la traduction ne pourrait pas être considéré comme un problème de cryptographie. Quand je regarde un article en russe, je me dis : « ceci a été écrit en anglais mais a été encodé avec des symboles étranges. Je vais maintenant procéder à son décodage ».*

(Weaver, lettre à Wiener du 4 mars 1947).

One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’

Les réserves de Wiener

- *En ce qui concerne la traduction automatique, j'ai peur que les frontières des mots dans les différentes langues soient trop vagues (…) pour rendre l'idée d'un système quasi automatique de traduction possible.*

(réponse de Wiener à Weaver, lettre du 30 avril 1947)

As to the problem of mechanical translation, I frankly am afraid that the boundaries of words in different languages are too vague (…) to make any quasi-mechanical translation scheme very hopeful.

Les systèmes à base de règles (1950-1990)
Une approche à la fois intuitive et naïve

- Deux éléments de base
  - Dictionnaires bilingues

    Dog ↔ *Chien*,

    Bank ↔ *Banque*, *Rive*…
  - Règles de transfert (traduction de mots ou groupes de mots en contexte)
    - *Bank* ↔ *Banque* si *argent, prêt…* dans le contexte
    - *I want him to do it.* ↔ *Je veux qu'il le fasse*. (**je veux lui faire cela*)

Triangle de Vauquois

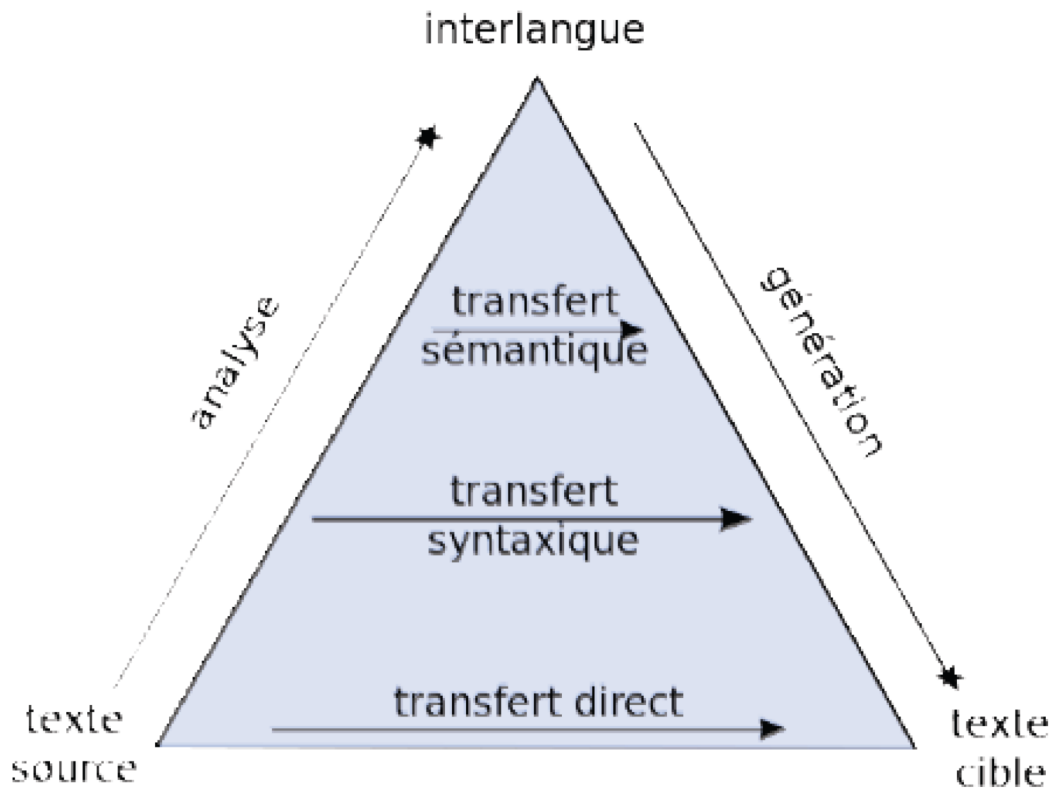


Schéma proposé par B.Vauquois à la fin des années 1960

Pourquoi ça fonctionne mal ?

- Impossibilité de prédire le sens de mots en contexte (car il y a une infinité de contextes possibles)
- Impossibilité de « couvrir » toute la langue (car il y a une infinité de phrases possibles)
- Lourdeur de mise en œuvre (maintenance de milliers de règles, contradiction entre règles)

  - Cf. « *The spirit is willing, but the flesh is weak* » traduit (anglais ⇌ russe) par « *The whisky is strong, but the meat is rotten* » (en fait, exemple apocryphe, cf. Hutchins)

La traduction statistique (1990-2014)

- Traiter la langue sur une base statistique
  - Codage / décodage de messages pendant la guerre
  - Succès de la transcription automatique de la parole
  - Phénomène statistiques en langue (cf. loi de Zipf, Pareto, procimité sémantique)

- Disponibilité de « gros corpus parallèles »
  - Trouver des équivalents lexicaux
  - Trouver des règles d'ordonnancement des mots (*voiture rouge* ↔ *red car*)
  - Cf. mémoires de traduction (*Linguee*, en ligne)

Un corpus parallèle : le Hansard

Texte français	Texte anglais
J'ai fait cette comparaison et je tiens à m'arrêter sur ce point.	I have looked at this and I want to talk about it for a second.
L'article 11 du projet de loi crée tellement d'exceptions qu'il va bien au-delà de l'article 21 de la convention, au point de carrément compromettre l'objet même de celle-ci.	Clause 11 in the bill creates so many exceptions that it goes well beyond article 21 of the treaty and basically completely undercuts the intention of the convention itself.
Je cite l'article 21 de la convention.	I will read what article 21 says.
C'est assez simple:	It is pretty straightforward:
Chaque État partie encourage les États non parties à la présente Convention à la ratifier, l'accepter, l'approuver ou y adhérer [...] Each State Party shall encourage States not party to this Convention to ratify, accept, approve or accede to this Convention	Chaque État notifie aux gouvernements de tous les États non parties à la présente Convention [...] Each State Party shall notify the governments of all States not party to this Convention

Modèles IBM

- Mis au point par une équipe d'IBM (1987-1993)
  - Equipe de F. Jelinek (Thomas J. Watson Research Centre)
  - Parallèle avec la transcription de la parole : signal audio ⇨ texte écrit
  - Traduction : langue source ⇨ langue cible
- Peut-on envisager une traduction (par transfert direct) à partir de corpus bilingues alignés ?

Représentation intuitive du modèle IBM1 (1/4)

Schéma d'après Koehn 2009

- Initialisation des alignements. Chaque mot anglais est relié à l'ensemble des mots français avec un lien équiprobable.

Représentation intuitive du modèle IBM1 (2/4)

Schéma d'après Koehn 2009

- Repérage des liens les plus fréquents
  - lien entre « la » et « the » dans l'exemple
- Renforcement de ces liens (au détriment des autres liens et des autres alignements possibles)

Représentation intuitive du modèle IBM1 (3/4)

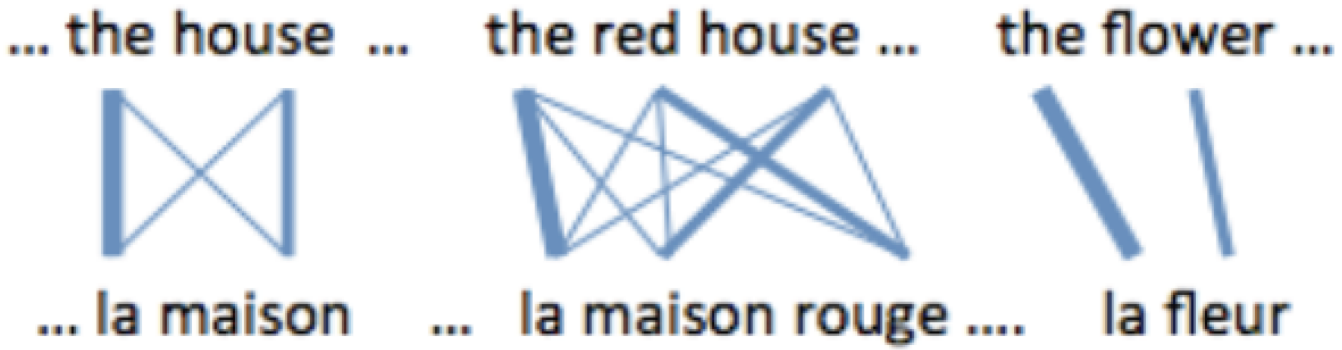


Schéma d'après Koehn 2009

- Identification des autres liens les plus probables
  - Maison ⇔ house », fleur ⇔ flower et red ⇔ rouge
- Renforcement de ces liens

Représentation intuitive du modèle IBM1 (4/4)

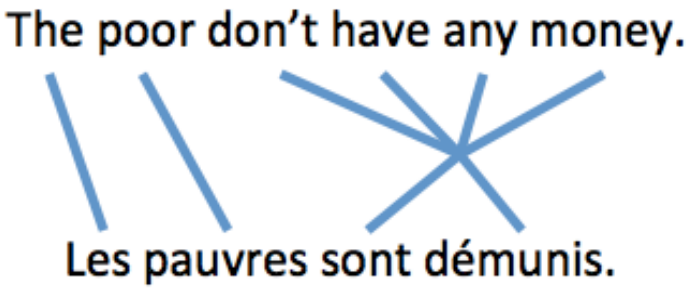
Schéma d'après Koehn 2009

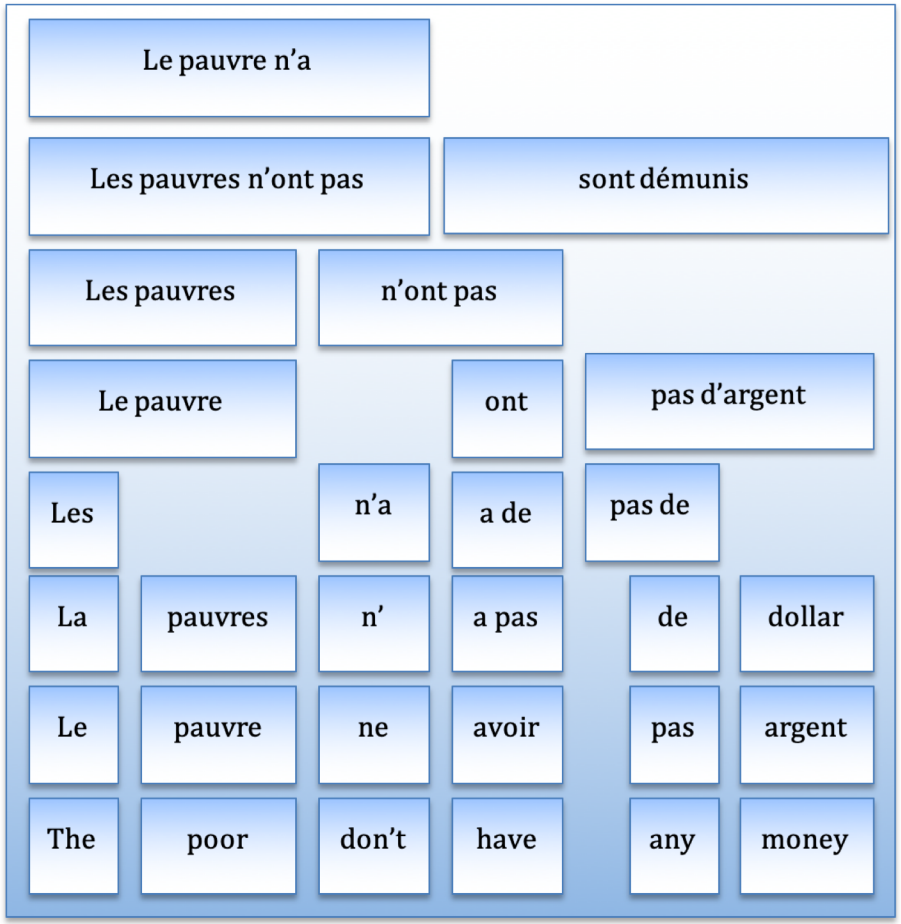
- Convergence vers une structure stable
- Les autres liens sont supprimés ou ont une probabilité très faible

Bilan sur les modèles statistiques

- Des modèles en apparence « simplistes »…
- … mais, de fait, si simplistes que ça ! (traduction par « segments »)
  - Séquences à trou, séquences avec chevauchement, etc.
  - Modèles syntaxiques, hiérarchiques…
  - Modèles de séquences ⇨ vers le transfert syntaxique
- Domination de modèles issus des travaux d'IBM
  - Bonnes performances lors des évaluations
  - Approche adoptée par Google et Microsoft

Traduction par segments





Source : Poibeau, Babel 2.0

Pourquoi ça fonctionne ? (…dans une certaine mesure…)

- Robustesse de ces modèles
  - Ils exploitent la richesse des données, dont la masse explose !
  - Ils couvrent en priorité les faits linguistiques les plus significatifs
  - Ils sont finalement beaucoup plus riches que des règles écrites à la main

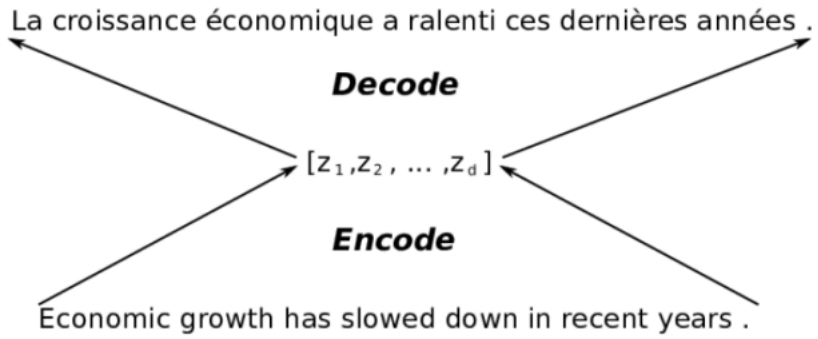
    « Every time I fire a linguist, the performance of our speech recognition system goes up. », Workshop on Evaluation of NLP Systems, Wayne, Pennsylvania, US, déc. 1988

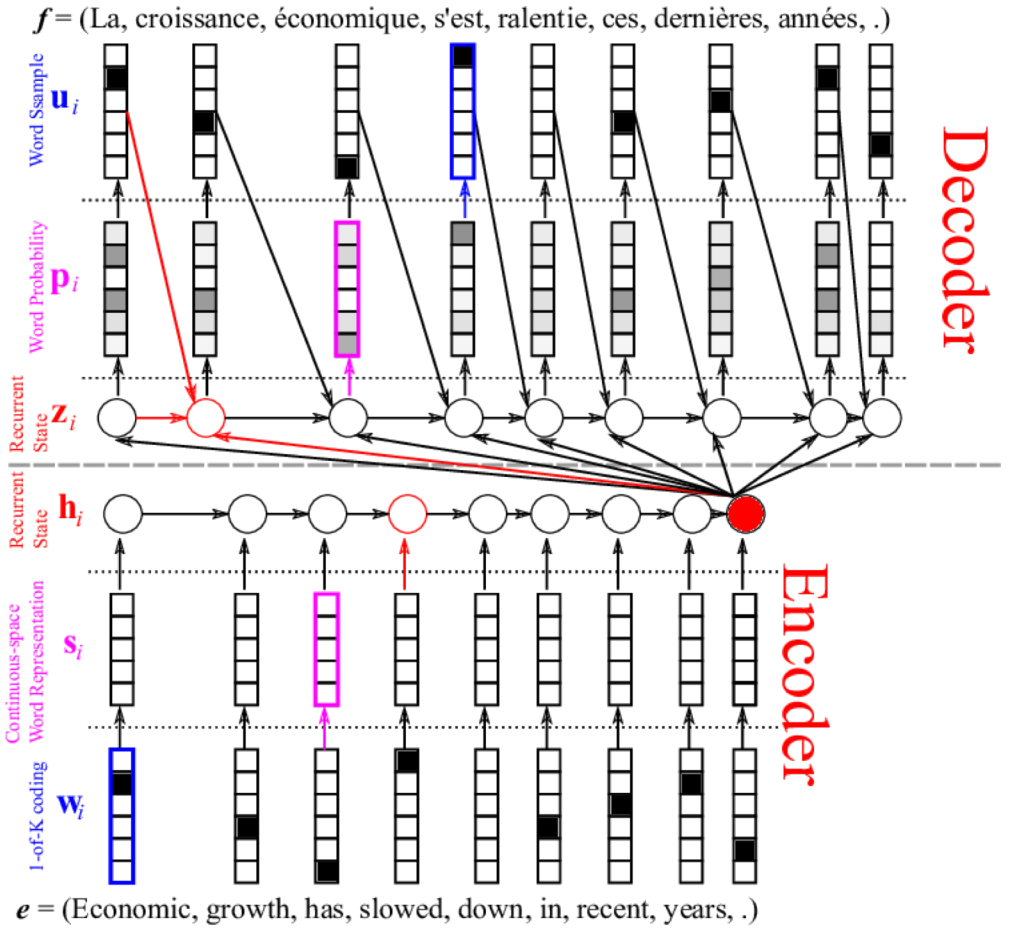
- Limites
  - L'approche reste locale : assemblage de fragments disparates
  - Nécessité de disposer de grandes masses de données

L’approche neuronale (2014-)

- Dans la lignée des approches statistiques (segments)
- Avantage des réseaux de neurones
  - Modélisation directe de la phrase (pas de segments de phrase à assembler)
  - Meilleure représentation du sens des mots
  - Meilleure représentation de la notion de contexte
- Simplicité et élégance des principes de codage
  - Décodage / encodage, cf. Weaver
  - (simplicité en partie remise en cause au fur et à mesure des avancées techniques)

Schématiquement…





Source : https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-2/

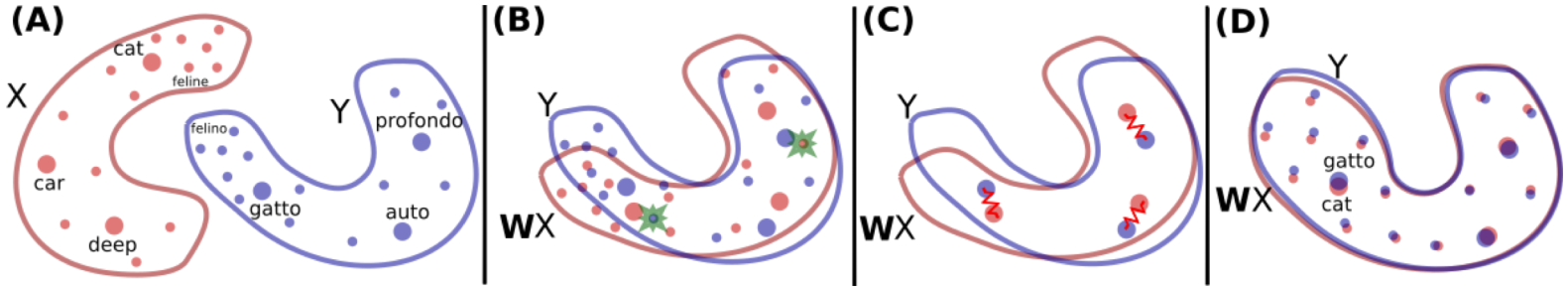
« Vectoriser » le sens des mots

- Encodage du sens des mots dans des vecteurs de taille fixe (typiquement, dimension 300 à 500). Chaque case représente un ensemble de contextes, c'est-à-dire une notion sémantique latente

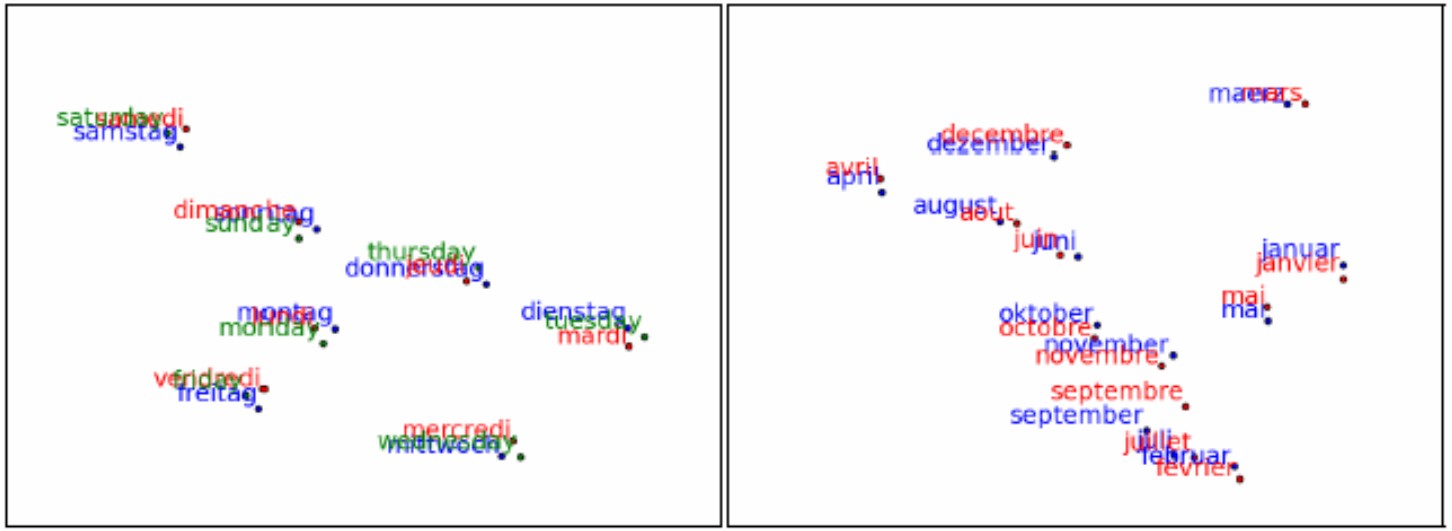
Source : Blog d'Adrian Acolier, *The morning paper* : https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/.

- Calcul sur les vecteurs : *King - Man + Woman = Queen* (Mikolov, 2013)

Représentations multilingues
(*Multilingual Embeddings*)



Source : projet MUSE (Multilingual Unsupervised and Supervised Embeddings) de l'équipe Facebook Paris https://github.com/facebookresearch/MUSE.



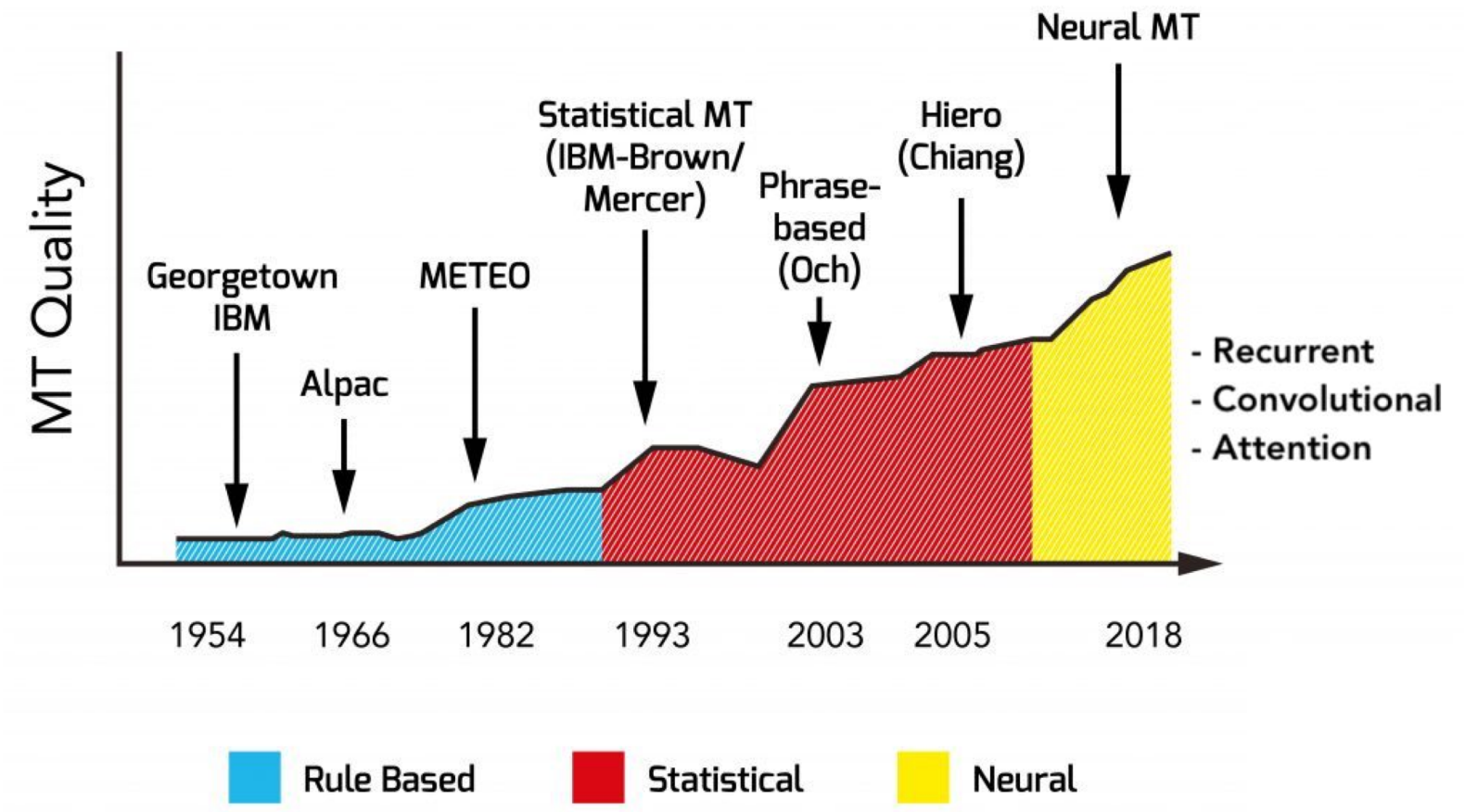
Source : Hermann and Blunsom, 2013. Cf. https://arxiv.org/abs/1312.6173

Et demain ?

- Un double enjeu
  - Mieux analyser les langues éloignées de l’anglais
  - Mieux analyser les langues « sous-dotées » (sans corpus parallèle)

- Des débuts de solution
  - Apprentissage d’un « espace sémantique »
  - Alignement d’espaces sémantiques
  - Traduction puis rétro-traduction et correction du biais observé

Evolution de la qualité de la TA



https://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf (revu par Iconic)

- Comment fonctionne la traduction automatique aujourd'hui ? Quelle est son histoire ?
- La TA est-elle aujourd'hui utilisable ? Comment ? Pour quoi faire ? Dans quels contextes ?
- Conclusion

Comment mesurer la qualité d'un système de TA ?

- Sujet de recherche en soi
  - Grande subjectivité de l'évaluation humaine
- Evaluation automatique
  - Obtenir la mesure la mieux corrélée avec une évaluation humaine
  - En pratique : mesure pondérée du nombre de séquences communes entre résultat automatique et traduction(s) humaine(s)
  - Pondération : longueur des séquences, etc. (n-grammes, en général : 4-gramme, cad séquences de 4 mots consécutifs)
  - Technique sommaire mais efficace pour l'anglais (mesure BLEU)

Etat des lieux

- La TA a aujourd'hui une qualité suffisante pour améliorer la productivité pour certaines langues (Koponen2016)

  MT can produce sufficient quality output today for commercial use with PE and has the potential to improve productivity in some languages

- La TA neuronale va se répandre de plus en plus dans les circuits de traduction en milieu professionnel

  Integrated into translation workflows in various settings & trend expected to grow stronger with neural MT

- Quel usage ? Dans quels milieux ? Pour quels types de documents ?

H. Martikainen, 2019

Une pratique déjà répandue

- TA déjà largement adoptée par les institutions internationales
  - ex. CE DGT & CdT (NMT *eTranslation*), WTO (SMT Moses), etc.
- Nombreuses d'études d'impact
  - ex. OCDE Translation Division (Champsaur 2019), Canadian Translation Bureau (avec l'Univ. de Montréal), La Poste suisse (Girletti 2019), etc.
- TA + post-édition, la clé du succès ?
  - Cf. agence de localisation internationales (Guerber of Arenas & Moorkens 2019)

H. Martikainen, 2019

Un exemple : la Commission européenne

EC DGT: plus de 2 M pages / an – 24 langues – 1 560 traducteurs

- *MT@EC*: statistique (2013) – *eTranslation*: TA neuronale (2017)
- Etude institutionnelle : besoins, compétences, limites et intégration / acceptation par les traducteurs en place
- Intégration de la TA dans le circuit de traduction : propositions de suggestions de traduction
- Evaluation d'usage (*MT@EC*): la post-édition est jugée acceptable pour la plupart des langues
- Bonne acceptation de la TA au sein de la DGT

Kluvanec 2017, Rossi & Chevrot 2019, H. Martikainen 2019

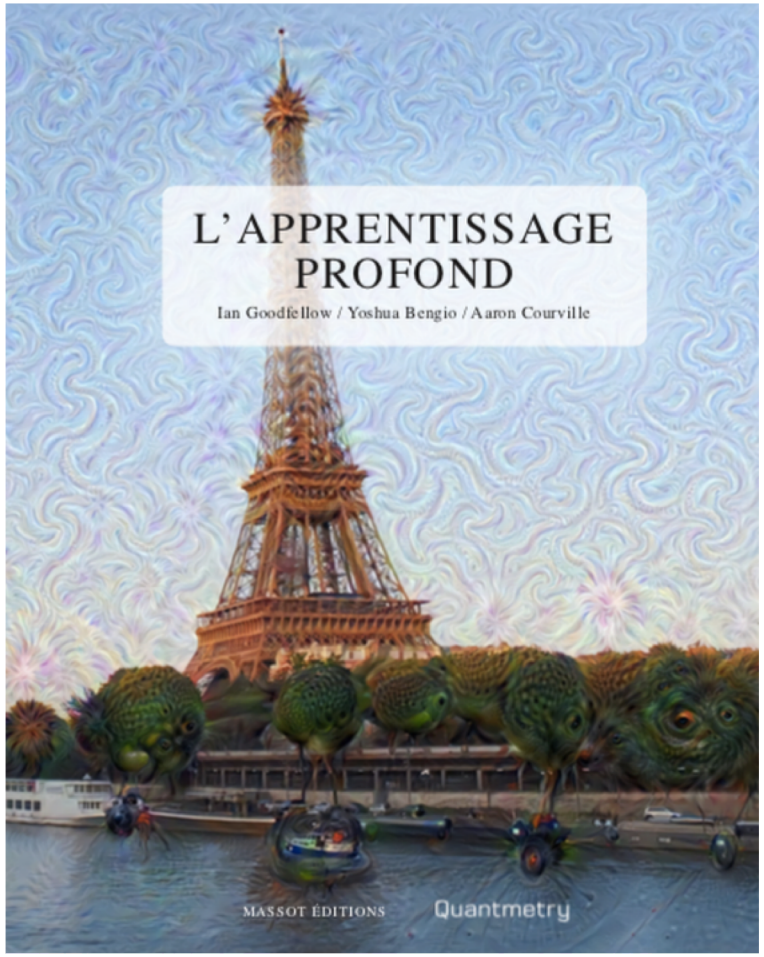
Pour quels types de documents ?

- La TA offre une traduction « littérale »
  - Pas de reformulation
  - Pas ou peu d'adaptation à la langue cible (expressions figées…)
  - Textes compréhensibles, mais encore peu idiomatiques

- Conséquences
  - Convient pour les documents techniques
  - Convient aux dépêches, au style journalistique
  - Ne convient pas aux textes littéraires
  - « Zone grise » pour laquelle la TA peut être utile, ou non

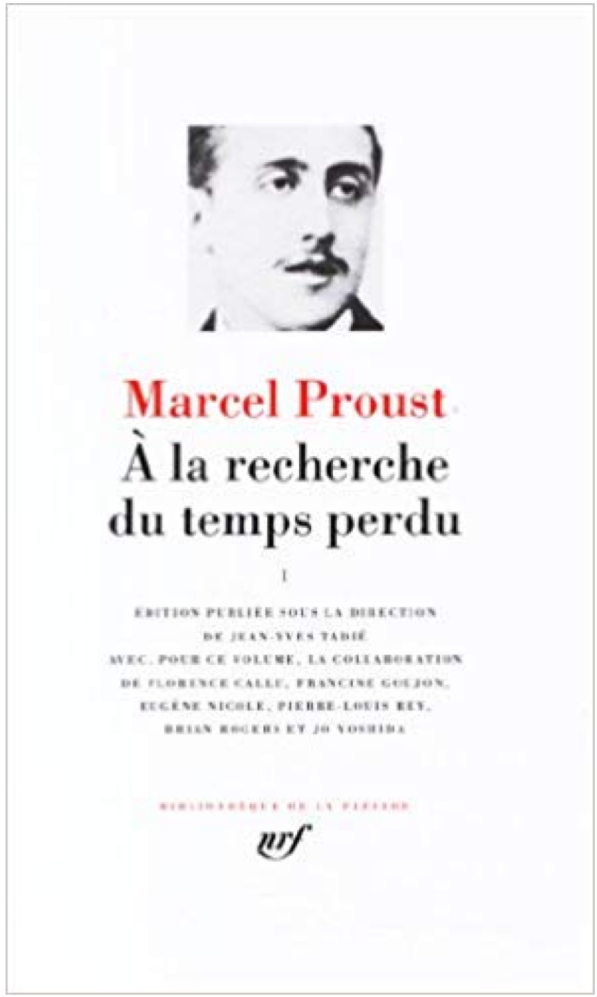
Traduire des livres automatiquement ?

En partie, oui…



Traduire des livres automatiquement ?

Mais non…



- Comment fonctionne la traduction automatique aujourd’hui ? Quelle est son histoire ?
- La TA est-elle aujourd’hui utilisable ? Comment ? Pour quoi faire ? Dans quels contextes ?
- Conclusion

En guise de synthèse…

- Facteurs à considérer
  - La qualité de la TA dépend des langues considérées
  - L'intérêt semble plus marqué pour les traducteurs moins expérimentés
  - Travail de préparation des données (structures des documents, termes techniques…)

- Risques de la TA en milieu professionnel
  - Traduction en apparence correct mais avec faux-sens, omission, etc.
  - Mise à l'écart du texte source
  - Mise en œuvre non immédiate

Les traducteurs sont-ils voués à être remplacés par la TA ?

- Non !
  - Problèmes de couverture langagière
  - Problème d'adaptabilité au domaine
  - Problèmes de fiabilité

- Sources de tension pour les traducteurs humains
  - Mécanisation du travail
  - Pression sur les délais
  - Nivellement des traductions, vers une traduction littérale
  - Pression sur les coûts

De nouvelles opportunités ?

- Identifier des documents à faire traduire par des traducteurs professionnels
- Traduire des documents qui ne seraient pas traduits sinon

- Au niveau de la recherche
  - Aborder la traduction avec peu de ressources (pas de corpus parallèle)
  - Soutenir des langues en danger

Merci de votre attention !