Université de Strasbourg

Traduction automatique
Introduction

Enzo Doyen

enzo.doyen@unistra.fr

2025 - LGC6KM41 - M2

Informations générales

- 6 séances en salle 4S04 le vendredi de **13 h** à **15 h** ; 26/09, 03/10, 10/10, 17/10 (**ou 24/10 ?**), 14/11, 21/11.
- Nécessité d'avoir un compte **Google Colab** ainsi qu'un compte **Hugging Face** à partir de la 3e séance pour les exercices pratiques et l'évaluation.

Objectifs

- Connaitre l'évolution historique des systèmes de traduction automatique (TA).
- Comprendre les concepts fondamentaux de la TA et le fonctionnement des différents types de systèmes de TA (modèles à base de règles, statistiques, neuronaux).
- Entrainer, sur la base de corpus parallèles, des systèmes de TA en utilisant les bibliothèques majeures (Keras, TensorFlow, Hugging Face).
- Mener une évaluation qualitative et quantitative de traductions automatiques, et interpréter des résultats d'évaluation de traduction automatique.

Évaluation

L'évaluation du cours se fait sur la base d'un **dossier à rendre** à la fin du semestre. Il se basera sur les éléments vus en cours et inclura un court rapport écrit ainsi que le code Python utilisé.

Plus d'informations seront données à la 3e séance.

🕮 Plan

I. Traduction automatique : définition, objectifs et défis

Définition de la TA

Définition : the process of using artificial intelligence to automatically translate text from one language to another without human involvement (https://aws.amazon.com/what-is/machine-translation/)

Définition de la TA

Définition : the process of using artificial intelligence to automatically translate text from one language to another without human involvement (https://aws.amazon.com/what-is/machine-translation/)

- Parole ?
- Taduction monolingue ?
- Degré d'intervention humaine ?

Objectifs de la TA

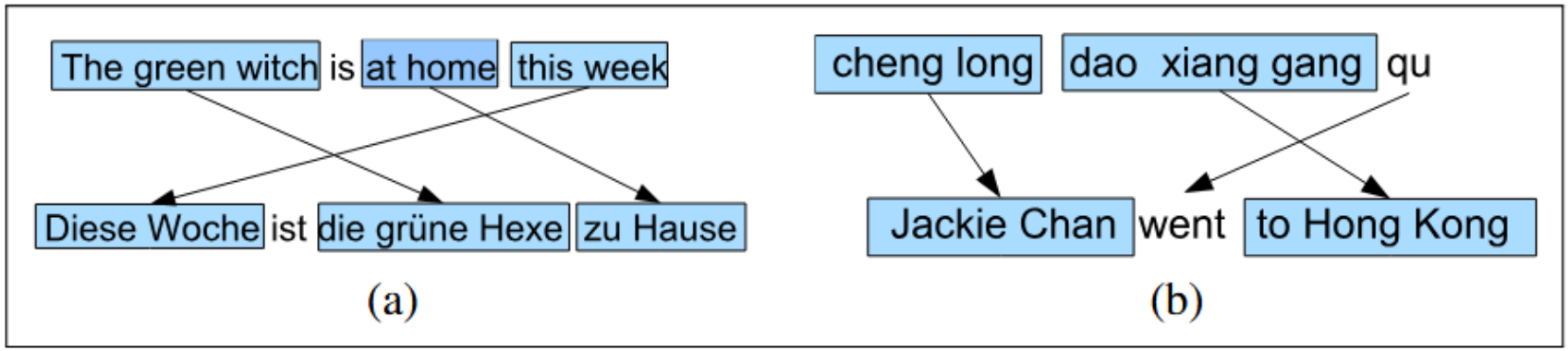
- passage d'un contenu dans une langue source à un texte dans une langue cible, tout en préservant le sens et le style du texte original ;
- permettre la communication rapide dans des contextes divers (professionnels, scientifiques, vie quotidienne…) ;
- adaptation à la culture cible ;
- mise à jour continue requise en phase avec l'évolution du langage.

Défis de la TA

Plusieurs défis relatifs à la TA, rapportés notamment par **Grass (2010)** :

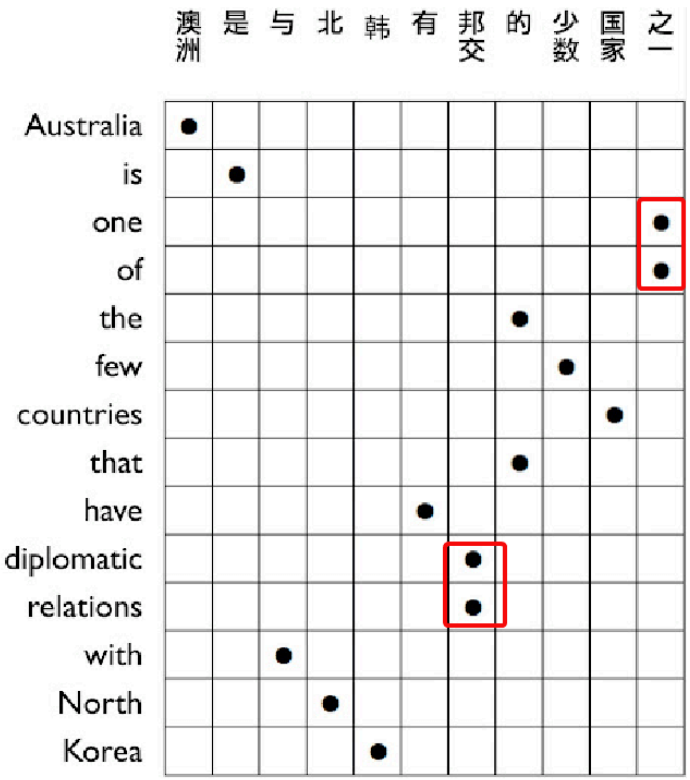
- ambigüité syntaxique, sémantique, référentielle
- contexte et nuances culturelles
- traduction créatives, sous contraintes
- expressions idiomatiques
- encodage grammatical différent : genre, nombre, temps…
- …

Défis de la TA : syntaxe



Source : **Jurafsky et Martin (2025)**

Défis de la TA : syntaxe



Source : **Li (2022)**

Défis de la TA : polysémie

« Ma sœur aime les **avocats**. »

Défis de la TA : polysémie

« Ma sœur aime les **avocats**. »

Fruit ou profession ?

Défis de la TA : ambigüité polysémique

« Il a perdu ses **fils**. »

Défis de la TA : ambigüité polysémique

« Il a perdu ses **fils**. »

Enfants ou bobines de fil ?

Défis de la TA : ambigüité syntaxique

Flying gliders can be dangerous.

Défis de la TA : ambigüité syntaxique

Flying gliders can be dangerous.

Défis de la TA : ambigüité syntaxique

Flying gliders can be dangerous.

Interprétation 1 : The act of flying gliders can be dangerous.
(= Piloter des planeurs peut être dangereux.)

« flying » gérondif, « gliders » nom objet

Interprétation 2 : Gliders that are flying can be dangerous.
(= Les planeurs en vol peuvent être dangereux.)

« flying » adjectif, « gliders » nom sujet

Défis de la TA : ambigüité référentielle

Paul a heurté le vase du pied et l'a cassé.

Défis de la TA : ambigüité référentielle

Paul a heurté le vase du pied et l'a cassé.

Paul a-t-il cassé le vase ou son pied ?

Défis de la TA : néologismes

Plusieurs types de néologismes **(Pinter et al., 2020)** :

- néologismes lexicaux (nouveaux concepts) : COVID long
- néologismes morphologiques (fusion de mots existants) : télétravail, vaccinodrome, *doomscrolling*, *binge-watching*
- néologismes sémantiques (nouvelle signification) : *ghosting*

Travaux de recherche récents intéressants sur les néologismes : **Lerner et Yvon (2025) ; Zheng et al. (2024)**

II. Histoire de la traduction automatique

Mémorandum de Warren Weaver (1949)

- Warren Weaver, mathématicien et cryptographe, a écrit un mémorandum influent qui a lancé les bases de la recherche en traduction automatique.

Mémorandum de Warren Weaver (1949)

- Warren Weaver, mathématicien et cryptographe, a écrit un mémorandum influent qui a lancé les bases de la recherche en traduction automatique.
- S'inscrit dans le contexte géopolitique de l'après-Second Guerre mondiale et de la Guerre froide, notamment avec la nécessité de **traduire des documents scientifiques et militaires**.

Mémorandum de Warren Weaver (1949)

La langue est vue comme **un code à décoder**, de manière similaire à la cryptographie :

> « It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the « Chinese code. » **If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?** »
>
> **– Weaver (1949)**

Mémorandum de Warren Weaver (1949)

Premières idées évoquées par Weaver :

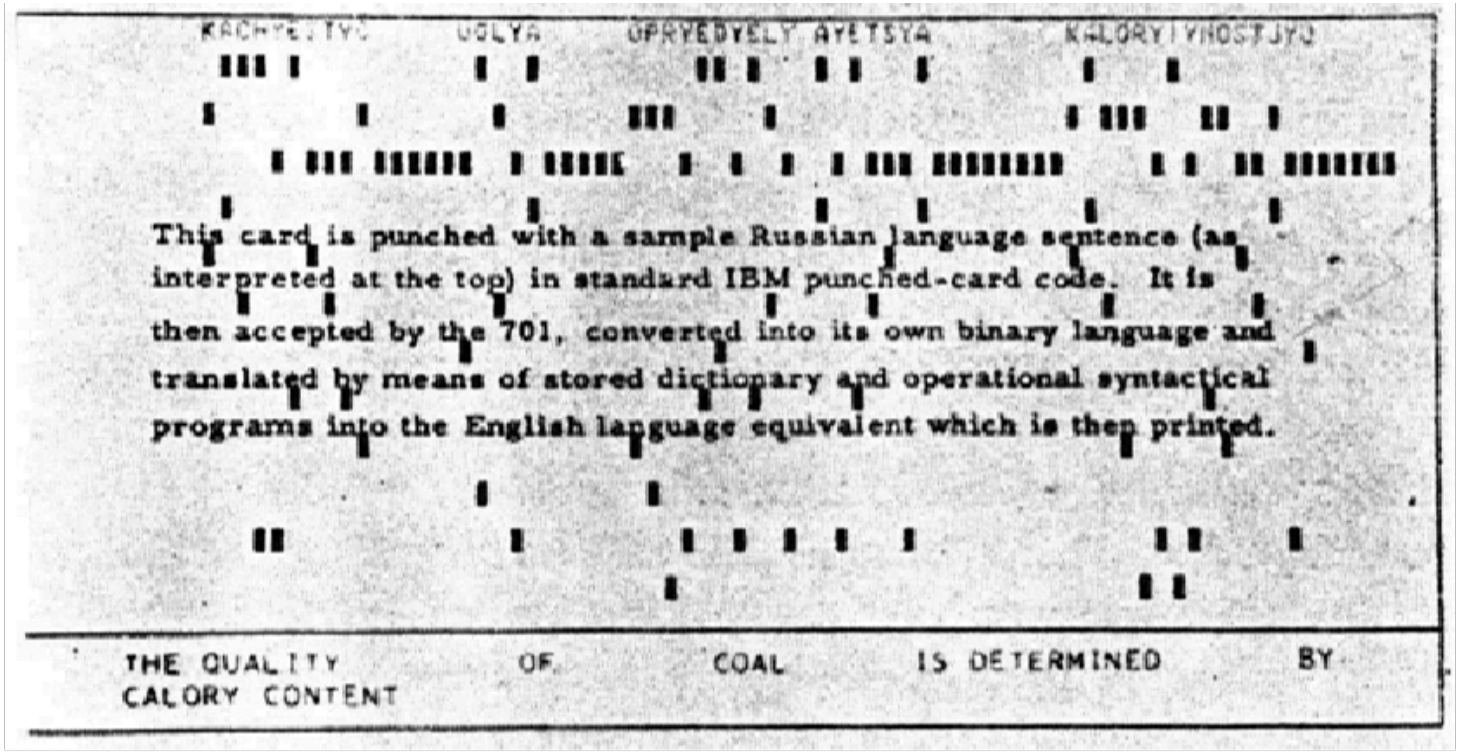
- traduction sur la base de dictionnaire ;
- désambigüisation par le contexte, en regardant les N mots voisins ;
- idée d'invariance linguistique.

> « Thus may it be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to **descend, from each language, down to the common base of human communication**. »
>
> **– Weaver (1949)**

Expérience Georgetown-IBM (1954)

- Première démonstration publique d'un système de traduction automatique.
- Système créé par l'Université de Georgetown et IBM pour traduire automatiquement des phrases du russe vers l'anglais.
- Utilisation d'un système simple avec un vocabulaire de 250 mots et 6 règles grammaticales.

Pour les détails techniques, voir **Hutchins (2004)**.


KACHYESTVO UGLYA OPREDYELYAYETSYA KALORIYNOSTJYU
This card is punched with a sample Russian language sentence (as interpreted at the top) in standard IBM punched-card code. It is then accepted by the 701, converted into its own binary language and translated by means of stored dictionary and operational syntactical programs into the English language equivalent which is then printed.
THE QUALITY OF COAL IS DETERMINED BY
CALORY CONTENT

Expérience Georgetown-IBM (1954)

Expérience jugée prometteuse et acclamée par la presse ; on dit que la traduction automatique serait un problème résolu dans seulement quelques années.

Exemples de titres de presse :

- "It's all done by machine"
- "Robot brain translates Russian into King's English"
- "The bilingual machine"
- "Polyglot brainchild"

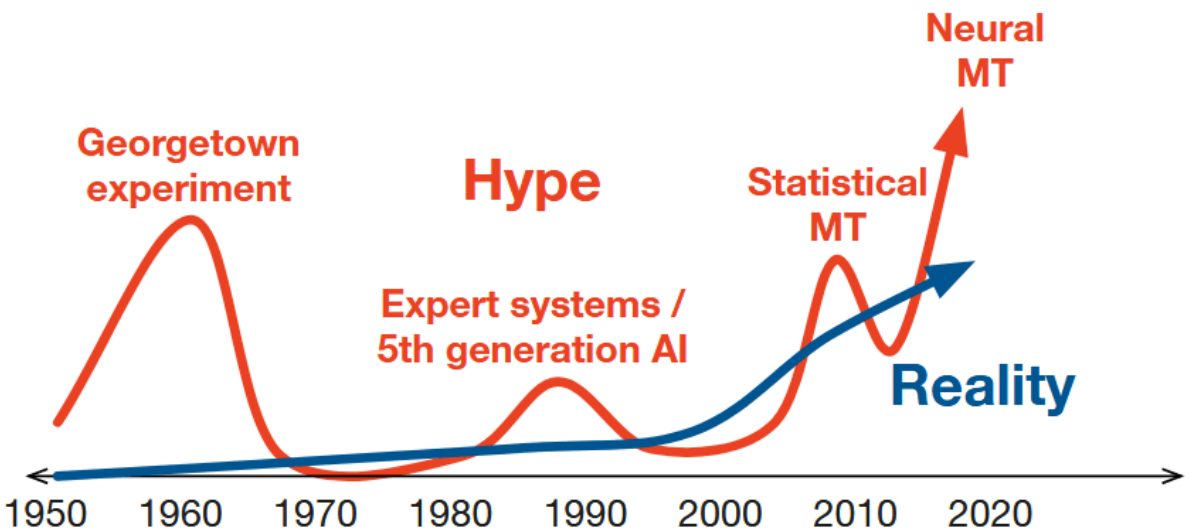


Figure 3.1 Machine translation (MT) hype cycle. For most of its history the expectations of machine translation quality have either over- or underestimated the real progress in quality, forming boom-and-bust cycles for research and development. Nevertheless machine translation has become a practical reality over the last 20 years. (Note: chart based on my subjective impressions, not hard facts.)

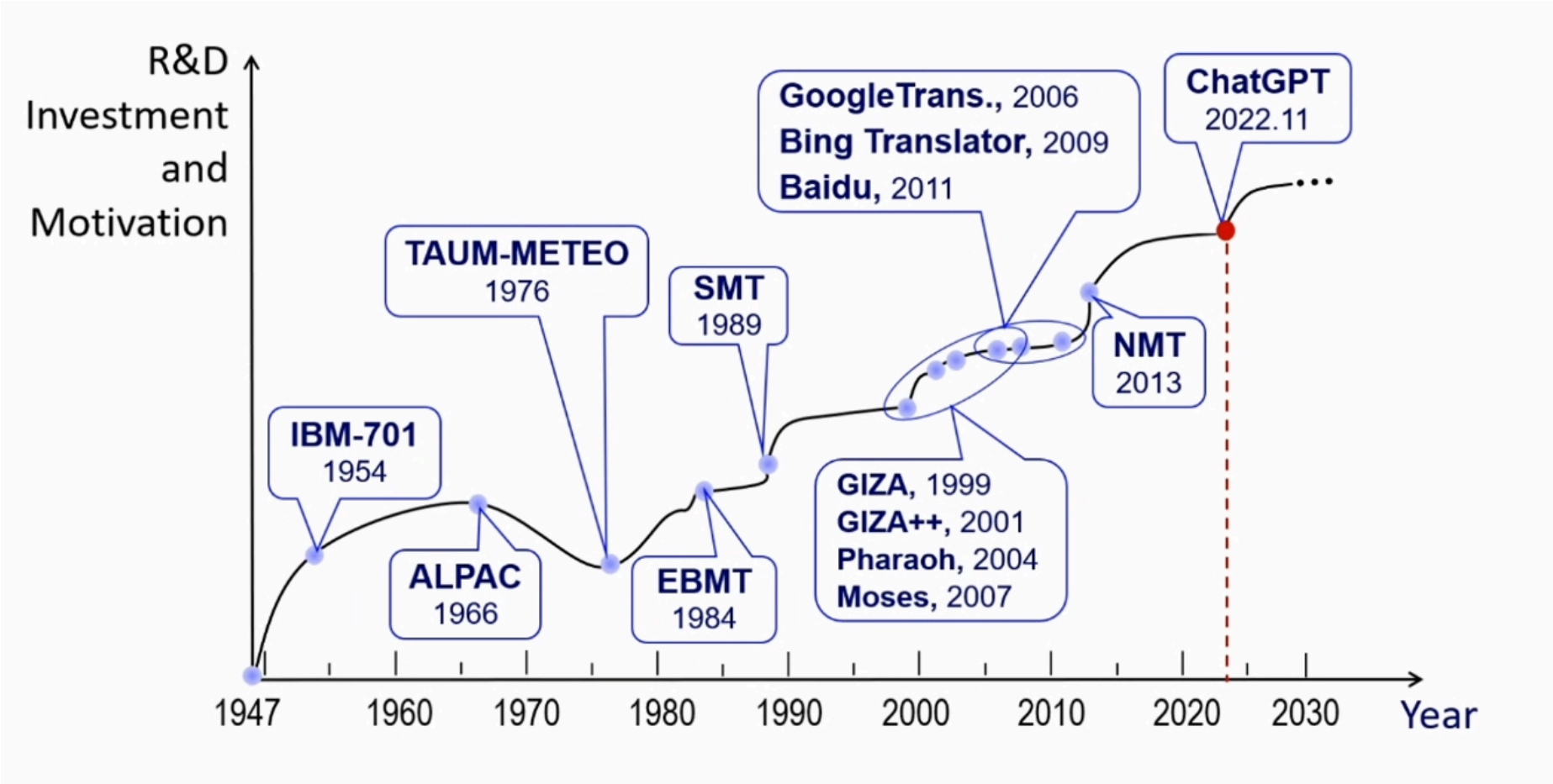
Koehn (2020)

Expérience Georgetown-IBM (1954)

En réalité…

- vocabulaire limité (250 mots) ;
- limité au domaine scientifique ;
- règles grammaticales simples (6 règles) et créées manuellement ;
- incapacité de prendre en compte la complexité et l'ambigüité du langage naturel.

Le système Georgetown-IBM était davantage une démonstration soigneusement organisée qu'un système de traduction généraliste et révolutionnaire.



Source : Chengqing Zong, *Presidential Address* (ACL 2025)

III. Premiers systèmes de TA : à base de règles et d'exemples

Traduction automatique à base de règles (RBMT)

Idée principale : la traduction automatique peut être réalisée en utilisant des règles linguistiques et un dictionnaire.

Procédure : elle impliquait des linguistes et des programmeurs et programmeuses qui créaient manuellement un vaste ensemble de règles pour les langues source et cible.

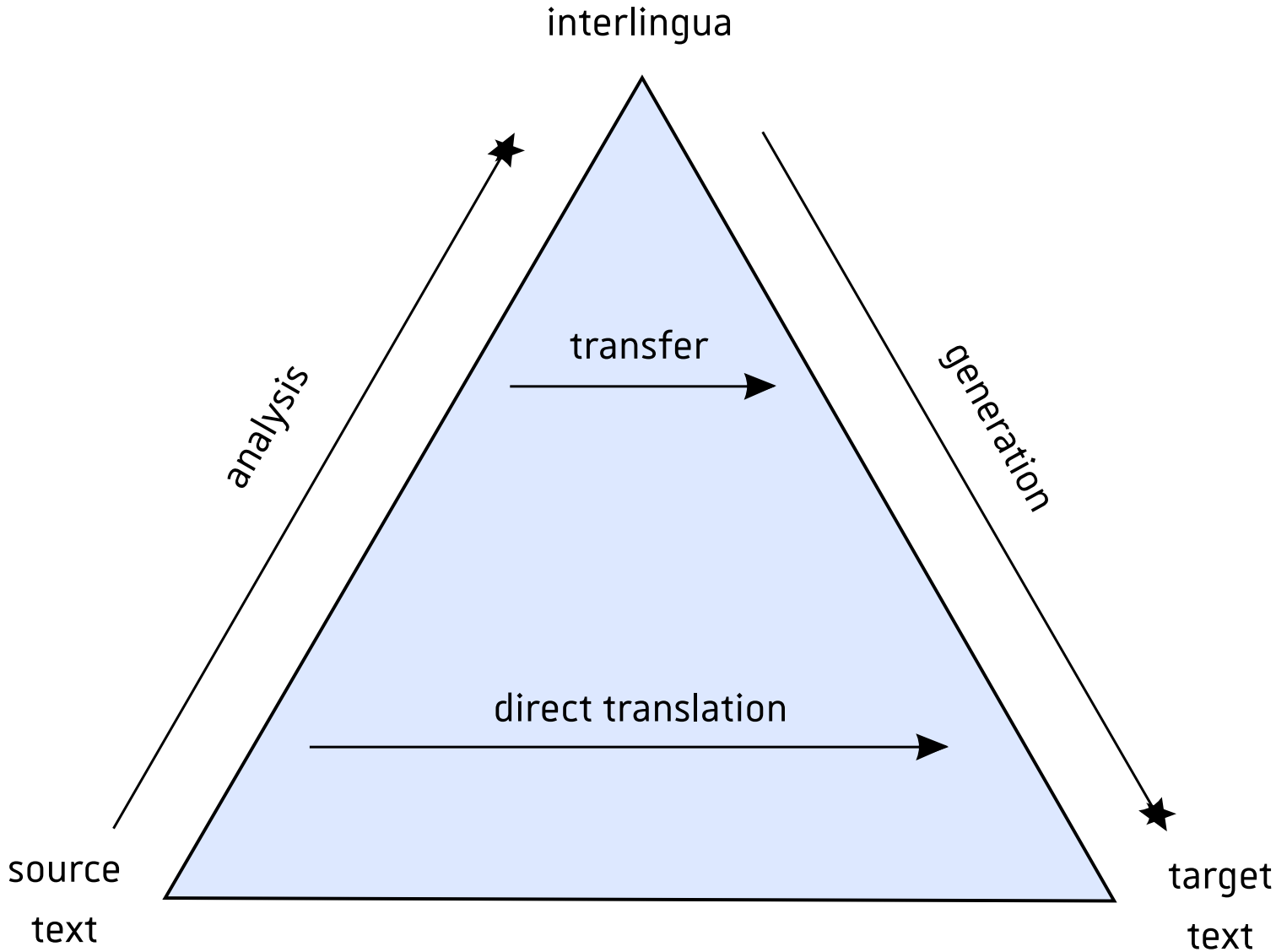
Traduction automatique à base de règles (RBMT)

Idée principale : la traduction automatique peut être réalisée en utilisant des règles linguistiques et un dictionnaire.

Procédure : elle impliquait des linguistes et des programmeurs et programmeuses qui créaient manuellement un vaste ensemble de règles pour les langues source et cible.

Types :

- **directe** : traduction plus ou moins mot à mot, avec des règles simples.
- **transfert** : analyse morphosyntaxique de la langue source, puis transformation vers la langue cible ; lexiques et grammaires.
- **interlingue** : traduction en une langue intermédiaire (langue pivot, p. ex. anglais pour Google Translate), puis vers la langue cible.


interlingua
analysis
transfer
generation
direct translation
source
text
target
text

Exemple de TA à base de règles : METEO

Exemple de système de traduction automatique à base de règles :

METEO (Chandioux, 1976)

- Initiative de l'Université de Montréal dans les années 1970 pour traduire les bulletins météorologiques de l'anglais vers le français.
- Traduction de 1 000 mots par minute, avec une charge d'environ 30 000 mots par jour.
- Première système dont les sorties sont rendues disponibles immédiatement au public (via la presse) sans intervention humaine.

Exemple de TA à base de règles : METEO

Quatre composants :

- dictionnaire de noms propres géographiques et de termes météorologiques ;
- dictionnaire de mots courants, avec leurs informations grammaticales ;
- analyseur syntaxique et morphologique pour identifier la structure des phrases en anglais ;
- générateur de phrases en français.

HIGH LEVEL

WOOD BUFFALO REGIONS

MOSTLY CLEAR AND COLD WITH PERIODS OF VERY LIGHT SNOW TODAY AND WEDNESDAY. HIGHS NEAR MINUS 10 BOTH DAYS. LOWS TONIGHT MINUS 20 TO MINUS 22.

HIGH LEVEL

WOOD BUFFALO

AUJOURD HUI ET MERCREDI GENERALEMENT CLAIR ET FROID AVEC TRES FAIBLES CHUTES DE NEIGE PASSAGERES. MAXIMUM POUR LES DEUX JOURS ENVIRON MOINS 10. MINIMUM CE SOIR MOINS 20 A MOINS 22.

Exemple de traduction anglais-français réalisée par METEO

Limites de la TA à base de règles

- effort manuel considérable pour créer et maintenir les règles, y compris les exceptions ;
- ressources linguistiques coûteuses et difficiles à créer ;
- difficulté à gérer les ambigüités du langage naturel
- uniquement applicables à des domaines très spécifiques avec des textes dont le contenu et la structure sont prévisibles.

Traduction automatique à base d'exemples (EBMT)

Idée principale : la traduction automatique peut être réalisée en utilisant des exemples de traductions précédentes.

Procédure : elle impliquait la création d'une base de données de phrases parallèles (corpus bilingue), qui servait de référence pour les nouvelles traductions.

TA à base d'exemples : les corpus

Distinction entre **corpus parallèle** et **corpus comparable** :

Définition : Un **corpus parallèle** est une collection de textes dans deux langues alignés au niveau du document, de la phrase ou des mots, permettant une comparaison et une traduction directe.

Définition : Un **corpus comparable** est une collection de textes dans deux langues qui ne sont pas nécessairement alignés, mais qui traitent de sujets similaires ou connexes, permettant une analyse comparative.

TA à base d'exemples : exemples de corpus parallèles

- **Europarl** : corpus de discours du Parlement européen, aligné au niveau des phrases, disponible dans 21 langues.
- **OpenSubtitles** : corpus de sous-titres de films et séries.
- **Hansard** : corpus de débats parlementaires canadiens (anglais-français).
- **OPUS** : collection de corpus parallèles multilingues, incluant des traductions de livres, de sites Web et d'autres documents.
- **Tatoeba** : corpus de phrases traduites dans de nombreuses langues, avec des exemples de phrases courtes et des traductions.

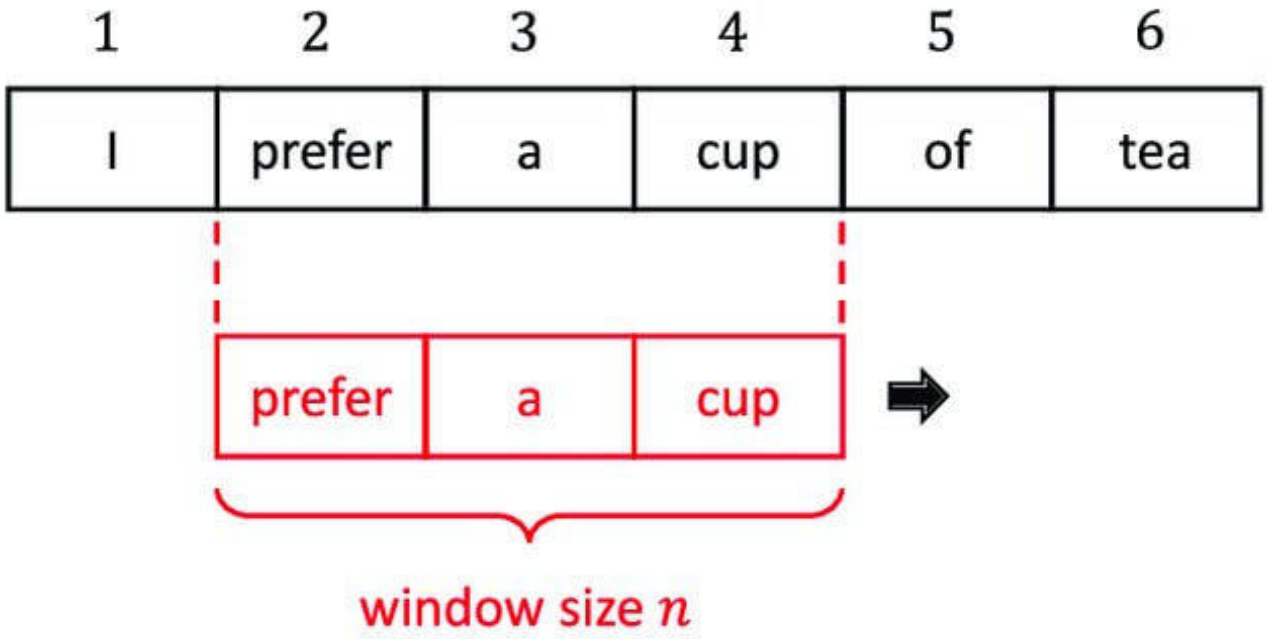
TA à base d'exemples : fonctionnement

La **TA à base d'exemples** utilise des techniques de recherche pour trouver des phrases similaires dans le corpus parallèle : le *matching* exploite les similarités entre les fragments source et cible par « analogie ». Il n'y a pas besoin d'une correspondance parfaite.

On peut par exemple comparer :

- les chaines de caractères ;
- les n-grammes (sous-chaines de longueur n) ;

TA à base d'exemples : n-grammes



Source : **Steur et Schwenker (2021)**

TA à base d'exemples : fonctionnement

La **TA à base d'exemples** utilise des techniques de recherche pour trouver des phrases similaires dans le corpus parallèle : le *matching* exploite les similarités entre les fragments source et cible par « analogie ». Il n'y a pas besoin d'une correspondance parfaite.

On peut par exemple comparer :

- les chaines de caractères ;
- les n-grammes (sous-chaines de longueur n) ;
- les séquences de catégories grammaticales ;
- les arbres syntaxiques.

TA à base d'exemples : fonctionnement

Traduction du syntagme japonais « N_1 no N_2 » en anglais **(Sumita et Iida, 1991)**

Japonais	Anglais
youka **no** gogo	the afternoon **of** the 8th
kaigi **no** mokuteki	the object **of** the conference
kaigi **no** sankaryou	the application fee **for** the **conference**
toukyou **deno** taiza	the stay **in** Tokyo
isshukan **no** kyuka	a week**'s** holiday
mittsu **no** hoteru	three **ø** hotels

TA à base d'exemples : fonctionnement

Traduction du syntagme japonais « N_1 no N_2 » en anglais **(Sumita et Iida, 1991)**

Japonais	Anglais
youka **no** gogo	the afternoon **of** the 8th
kaigi **no** mokuteki	the object **of** the conference
kaigi **no** sankaryou	the application fee **for** the **conference**
toukyou **deno** taiza	the stay **in** Tokyo
isshukan **no** kyuka	a week**'s** holiday
mittsu **no** hoteru	three **ø** hotels

Phrase à traduire : kyouto **deno kaigi**

Traduction proposée : the **conference in** Kyoto

TA à base d'exemples : limites

- nécessite un corpus parallèle de grande taille et de qualité ;
- manque de couverture dans les données d'exemple ; le contexte n'est pas toujours identique.

IV. Traduction automatique statistique

Traduction automatique statistique (SMT)

Idée principale : la traduction automatique peut être réalisée en utilisant des modèles statistiques basés sur des données d'exemple.

Procédure : elle implique l'utilisation de modèles statistiques pour apprendre les relations entre les langues à partir de corpus parallèles **alignés**.

TA statistique : alignement

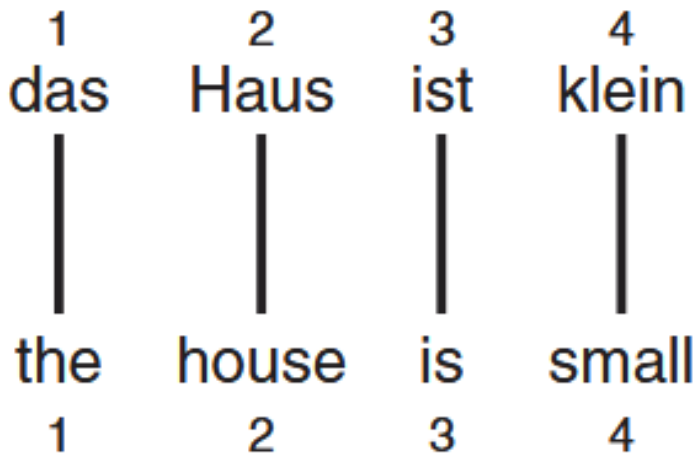
EN	FR
Steam Deck OLED has 30-50% more battery life. We fit a bigger battery into the case, and the OLED display draws less power. Combined with the updated, more efficient AMD APU, you have way more time to play your favorites.	L'autonomie de Steam Deck OLED est 30 à 50 % supérieure à celle du modèle LCD , ce grâce à une plus grande batterie et à l'écran OLED, qui est moins énergivore. Ajoutez à cela un nouvel APU d'AMD plus efficace, et vous obtenez encore plus de temps pour jouer à vos jeux favoris.

TA statistique : alignement

EN	FR
<s>Steam Deck OLED has 30-50% more battery life.</s> <s>We fit a bigger battery into the case, and the OLED display draws less power.</s> <s>Combined with the updated, more efficient AMD APU, you have way more time to play your favorites.</s>	<s>L'autonomie de Steam Deck OLED est 30 à 50 % supérieure à celle du modèle LCD</s> <s>, ce grâce à une plus grande batterie et à l'écran OLED, qui est moins énergivore.</s> <s>Ajoutez à cela un nouvel APU d'AMD plus efficace, et vous obtenez encore plus de temps pour jouer à vos jeux favoris.</s>

TA statistique : alignement au niveau des mots

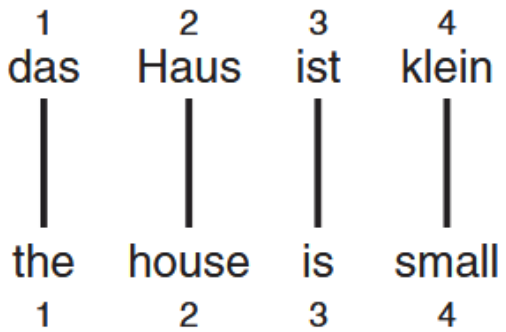
Les modèles de TA statistique utilisent l'alignement au niveau des mots pour établir des correspondances entre les mots de la langue source et ceux de la langue cible.



TA statistique : alignement au niveau des mots

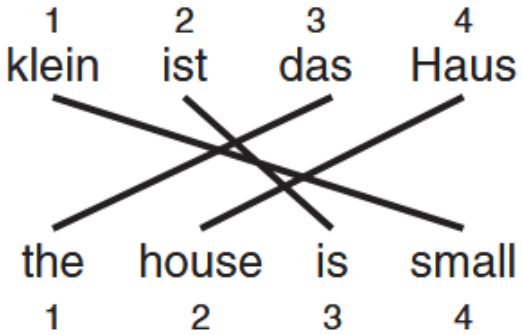
On peut définir une fonction d'alignement a qui associe chaque mot allemand j à un mot anglais i.

$$a : j \rightarrow i$$



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$$

TA statistique : alignement au niveau des mots



$$a : \{1 \to 3, 2 \to 4, 3 \to 2, 4 \to 1\}$$

TA statistique : approches

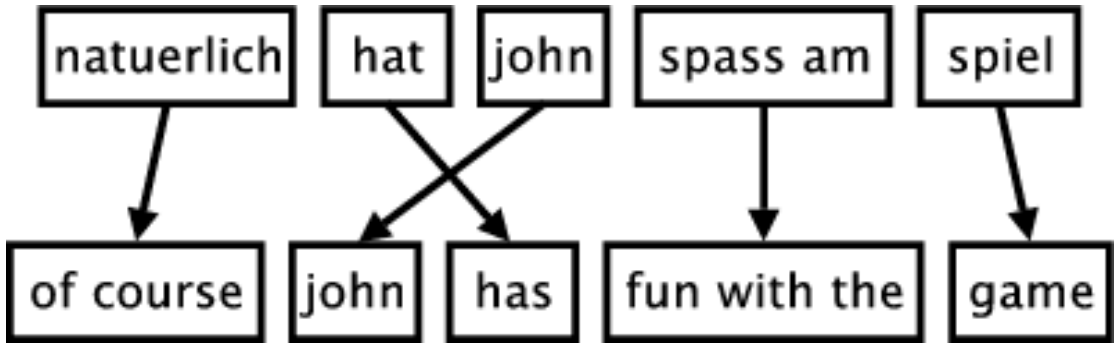
Plusieurs approches existent pour la TA statistique :

- **word-based** : la traduction se fait au niveau des mots individuels (cf. **GIZA++** et diapos précédentes) ;
- **phrase-based** : la traduction se fait au niveau de segments (cf. **Moses**) ;

TA statistique : approches

Plusieurs approches existent pour la TA statistique :

- **word-based** : la traduction se fait au niveau des mots individuels (cf. **GIZA++** et diapos précédentes) ;
- **phrase-based** : la traduction se fait au niveau de segments (cf. **Moses**) ;



TA statistique : approches

Plusieurs approches existent pour la TA statistique :

- **word-based** : la traduction se fait au niveau des mots individuels (cf. **GIZA++**) ;
- **phrase-based** : la traduction se fait au niveau de segments (cf. **Moses**) ;
- **syntax-based** : la traduction se fait en utilisant des structures syntaxiques, en tenant compte de la syntaxe des phrases source et cible ;
- **tree-based** : la traduction se fait en utilisant des structures hiérarchiques, en tenant compte des relations entre les phrases et les mots.

TA statistique : modèles

Les différentes approches de TA statistique se basent sur deux modèles :

- un **modèle de traduction**, chargé de déterminer les paires de phrases source et cible les plus probables à partir de corpus parallèles ;
- un **modèle de langue**, chargé de déterminer la probabilité d'une phrase cible donnée, en tenant compte de la structure grammaticale et du vocabulaire de la langue cible.

TA statistique : modèles

Étant donné une phrase source s et une phrase cible t, on cherche à maximiser la probabilité de la phrase cible t étant donné la phrase source s.

On peut reformuler ce problème avec le théorème de Bayes :

$$P(t|s) = \frac{P(t)P(s|t)}{P(s)}$$

où $P(t)$ est la probabilité de la phrase cible t, $P(s|t)$ est la probabilité de la phrase source s étant donné la phrase cible t, et $P(s)$ est la probabilité de la phrase source s (constante).

$P(t)$ vient du modèle de langue, et $P(s|t)$ du modèle de traduction.

TA statistique : modèle de langue

Modèle utilisé pour la langue cible, qui permet de déterminer la probabilité d'une phrase cible donnée ; essentiel pour évaluer la « fluency » de la traduction.

Ce type de modèle est souvent basé sur des n-grammes, et est entrainé sur un corpus monolingue de la langue cible. Il peut éventuellement être enrichi avec des informations syntaxiques ou sémantiques.

TA statistique : modèle de traduction

Modèle qui permet de déterminer la probabilité d'une phrase source donnée pour une phrase cible donnée ; évalue l'« adequacy » de la traduction.

Ce type de modèle est obtenu par une analyse statistique d'un corpus parallèle (probabilités des paires de séquences s et t), et peut être basé sur des alignements au niveau des mots ou des phrases.

TA statistique : exemples (*word-based*)

Source : « casa » | **candidats** : « house », « home », « building »

Modèle de langue $P(t)$:

- $P(\text{house}) = 0.01$
- $P(\text{home}) = 0.003$
- $P(\text{building}) = 0.005$

Modèle de traduction $P(s|t)$:

- $P(\text{casa}|\text{house}) = 0.8$
- $P(\text{casa}|\text{home}) = 0.6$
- $P(\text{casa}|\text{building}) = 0.1$

TA statistique : exemples (*word-based*)

Source : « casa » | **candidats** : « house », « home », « building »

Scores ($P(t) \times P(s|t)$) :

« house » : 0.8 × 0.01 = 0.008

« home » : 0.6 × 0.008 = 0.0048

« building » : 0.1 × 0.005 = 0.0005

« house » est le candidat le plus probable.

TA statistique : avantages et limites

Avantages

- Résolution des problèmes d'ambigüité syntaxique et sémantique grâce au recours au contexte local des mots.
- Capacité à gérer des langues avec des structures syntaxiques différentes.

Limites

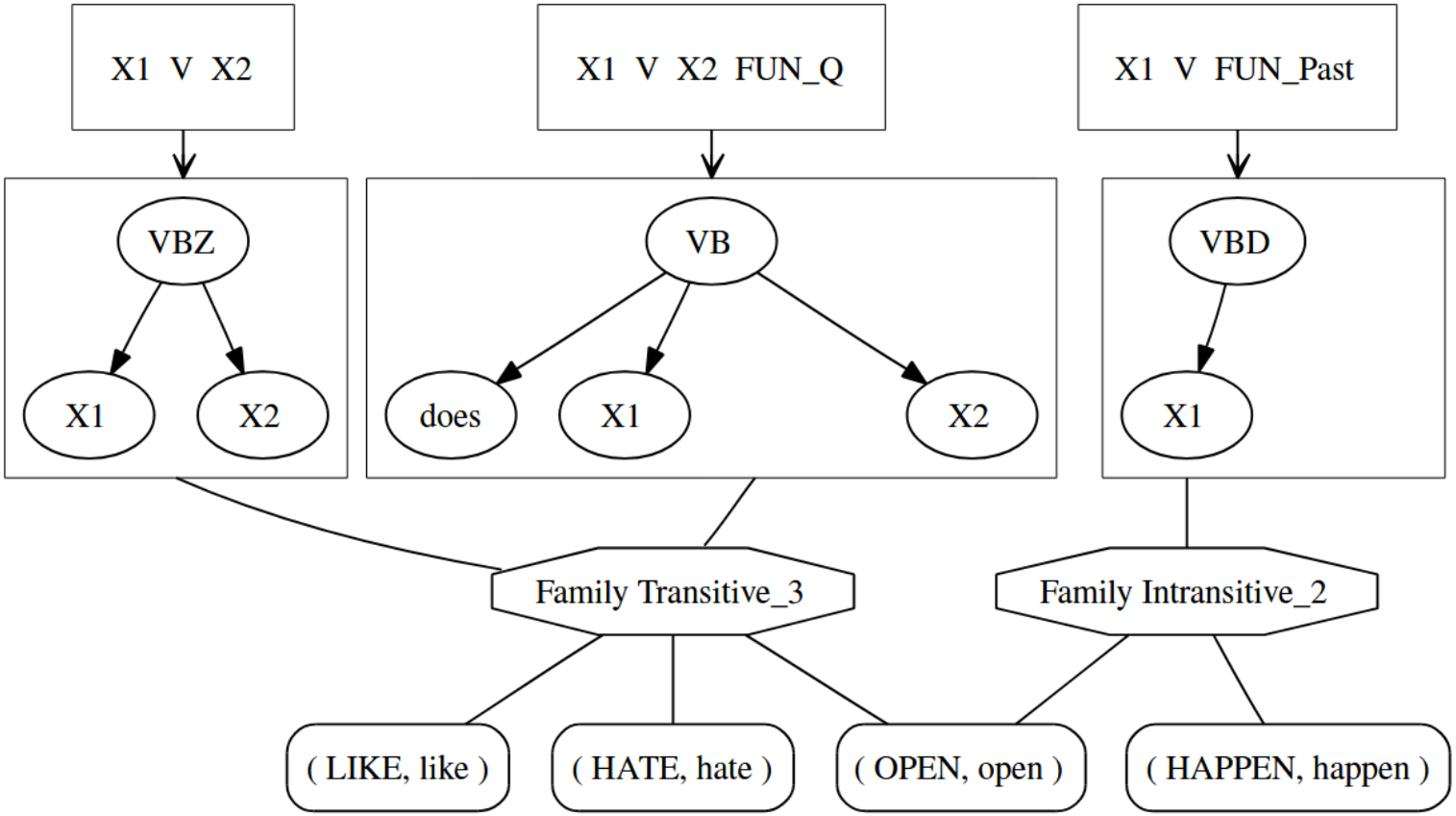
- Traitement limité à des séquences constituées de mots contigus.
- Nécessité de corpus parallèles de grande taille et de qualité pour l'entrainement des modèles.
- Simple mise en correspondance des séquences, et manque de prise en compte des informations linguistiques, ce qui donne parfois des traductions peu naturelles ou incohérentes.

TA statistique : avantages et limites

Limites

- Simple mise en correspondance des séquences, et manque de prise en compte des informations linguistiques, ce qui donne parfois des traductions peu naturelles ou incohérentes.

Problème résolu par les **systèmes factorisés**.



Source : **Shen et al. (2010)**

TA statistique : systèmes factorisés

- Extension des systèmes à base de séquences par l'utilisation d'informations linguistiques associées aux mots.
- Exploitation de corpus alignés et **annotés** avec des informations linguistiques (étiquettes morphosyntaxiques, lemmes, constituants syntaxiques, etc.).

TA statistique : systèmes factorisés, exemple

Soit la forme à traduire **apple** et la cible **pomme**.

TA statistique : systèmes factorisés, exemple

Soit la forme à traduire **apple** et la cible **pomme**.

1. **Traduction** : mise en correspondance des lemmes
   - apple → {**pomme**, Apple}

TA statistique : systèmes factorisés, exemple

Soit la forme à traduire **apple** et la cible **pomme**.

1. **Traduction** : mise en correspondance des lemmes
   - apple → {**pomme**, Apple}
2. **Traduction** : mise en correspondance des étiquettes de POS
   - NOUN → {**NOUN**, PROPN}

TA statistique : systèmes factorisés, exemple

Soit la forme à traduire **apple** et la cible **pomme**.

1. **Traduction** : mise en correspondance des lemmes
   - apple → {**pomme**, Apple}
2. **Traduction** : mise en correspondance des étiquettes de POS
   - NOUN → {**NOUN**, PROPN}
3. **Traduction** : mise en correspondance des étiquettes morphosyntaxiques
   - nc-s → {**ncfs**, ncms, ncfp, ncmp}

TA statistique : systèmes factorisés, exemple

Soit la forme à traduire **apple** et la cible **pomme**.

1. **Traduction** : mise en correspondance des lemmes
   - apple → {**pomme**, Apple}
2. **Traduction** : mise en correspondance des étiquettes de POS
   - NOUN → {**NOUN**, PROPN}
3. **Traduction** : mise en correspondance des étiquettes morphosyntaxiques
   - nc-s → {**ncfs**, ncms, ncfp, ncmp}
4. **Génération** : mot-forme généré selon les sorties des étapes précédentes
   - **apple** | **NOUN** | **ncfs** → **pomme**

V. Traduction automatique neuronale

Traduction automatique neuronale (NMT)

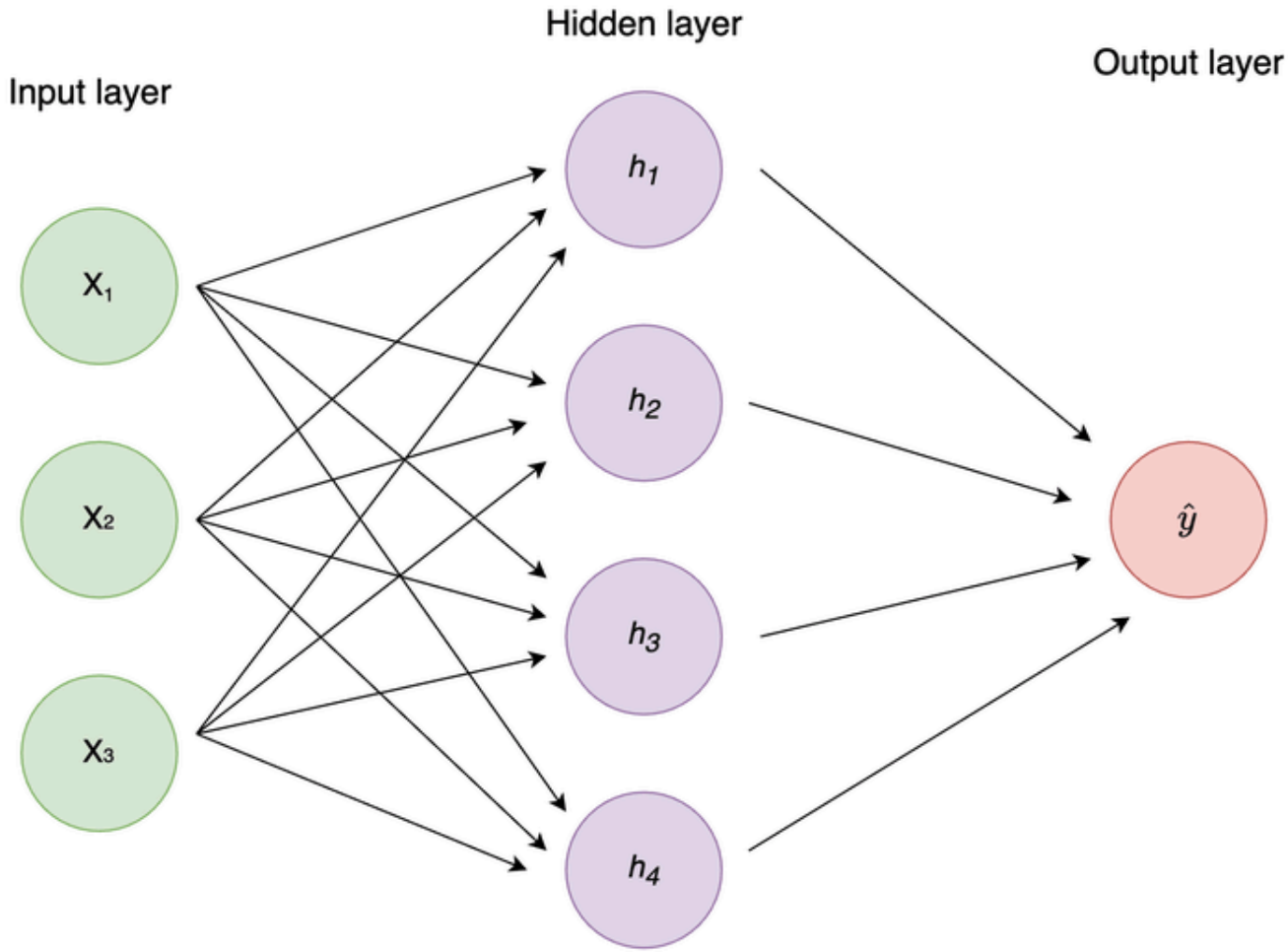
Approche en TA déjà explorée dès les années 90, mais la puissance de calcul de l'époque était insuffisante pour une application pratique en TA.

Montée en puissance de calcul et disponibilité de données massives ont permis le développement de la traduction automatique neuronale à partir des années 2010.

Traduction automatique neuronale (NMT)

Idée principale : la traduction automatique peut être réalisée en utilisant des réseaux de neurones profonds pour modéliser les relations entre les langues.

Procédure : elle implique l'utilisation de modèles de réseaux de neurones, entrainés sur de grands corpus parallèles.



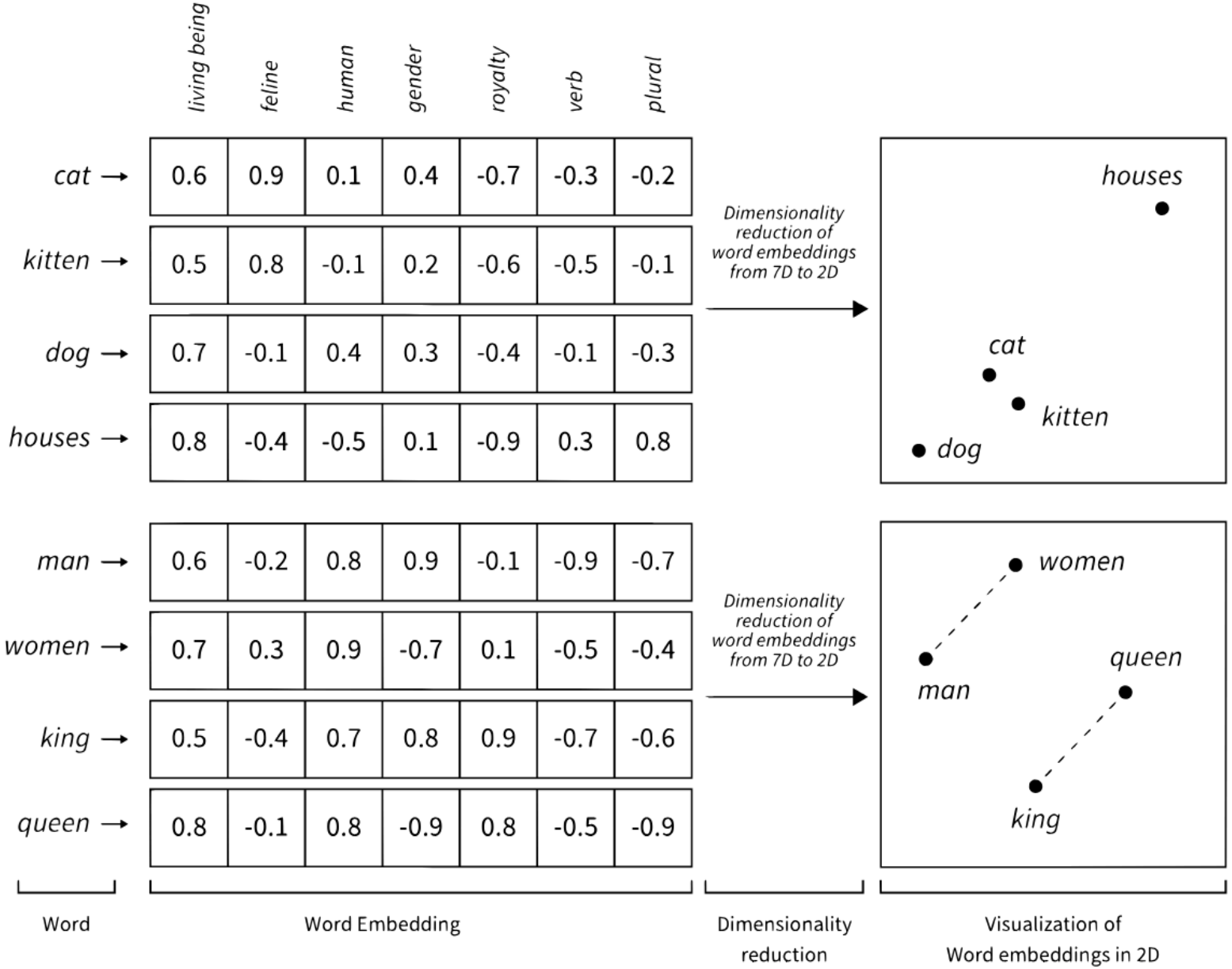
Source : Lasse Hassen

Traduction automatique neuronale (NMT)

- Réseaux de neurones : plusieurs « couches » de neurones interconnectés
- Chaque neurone reçoit des entrées (vecteurs/matrices), effectue un calcul et produit une sortie (vecteur/matrice)
- Les poids des connexions entre les neurones sont ajustés pendant l'entrainement pour minimiser l'erreur de prédiction
- Moyen très riche de représenter les informations et de les généraliser à partir du corpus d'entrainement

Traduction automatique neuronale (NMT)

En TAL neuronal, l'unité de base du traitement n'est plus le token, mais l'**embedding** (ou **plongement lexical**) : un vecteur dense qui représente le sens d'un mot dans un espace vectoriel selon ses contextes d'apparition.


living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 -0.7 -0.3 -0.2
kitten → 0.5 0.8 -0.1 0.2 -0.6 -0.5 -0.1
dog → 0.7 -0.1 0.4 0.3 -0.4 -0.1 -0.3
houses → 0.8 -0.4 -0.5 0.1 -0.9 0.3 0.8
man → 0.6 -0.2 0.8 0.9 -0.1 -0.9 -0.7
women → 0.7 0.3 0.9 -0.7 0.1 -0.5 -0.4
king → 0.5 -0.4 0.7 0.8 0.9 -0.7 -0.6
queen → 0.8 -0.1 0.8 -0.9 0.8 -0.5 -0.9
Dimensionality reduction of word embeddings from 7D to 2D
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
women
queen
man
king
Word
Word Embedding
Dimensionality reduction
Visualization of Word embeddings in 2D

Traduction automatique neuronale (NMT)

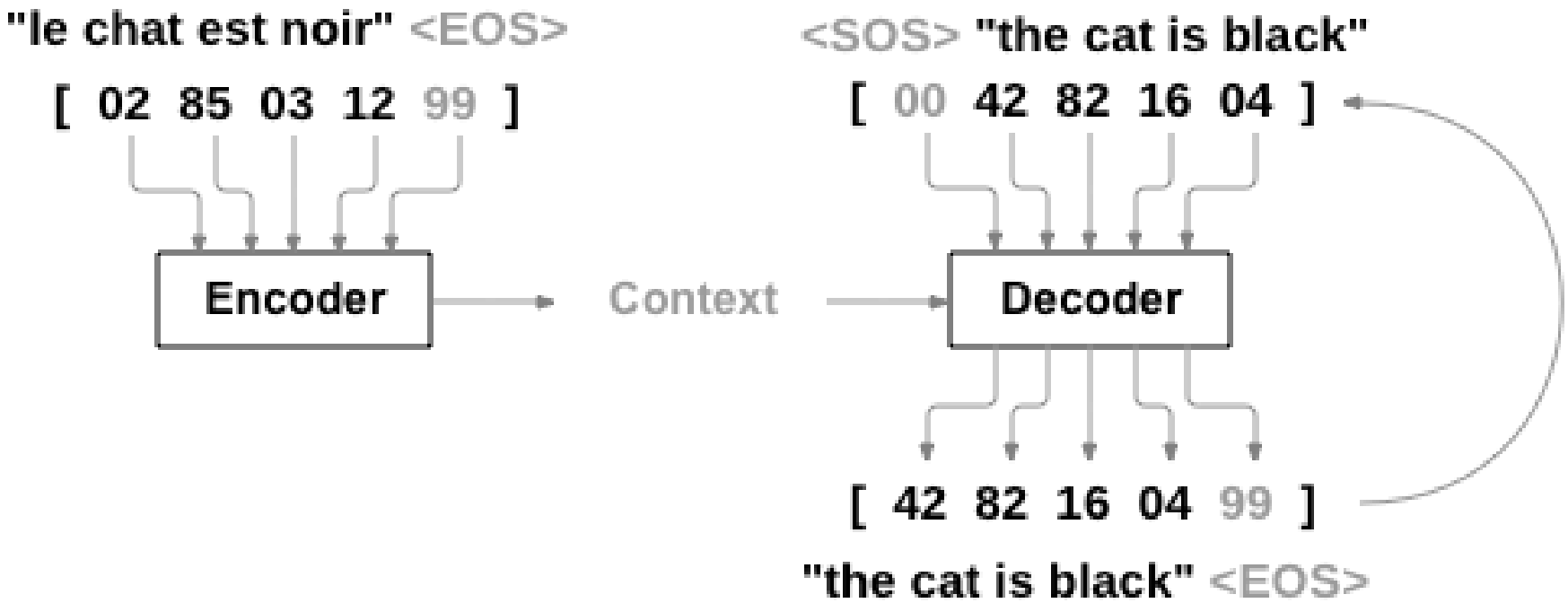
En TAL neuronal, l'unité de base du traitement n'est plus le token, mais l'**embedding** (ou **plongement lexical**) : un vecteur dense qui représente le sens d'un mot dans un espace vectoriel selon ses contextes d'apparition.

La TA neuronale permet d'utiliser la phrase au complet comme contexte (tandis que l'on exploite des fragments d'une longueur fixe en SMT).

Traduction automatique neuronale (NMT) : architecture encodeur-décodeur

Les modèles de traduction automatique neuronale sont souvent basés sur une **architecture encodeur-décodeur** (aussi appelée **séquence-à-séquence**), qui comprend trois composants :

- **encodeur** : transforme la phrase source en une représentation vectorielle (embedding) qui capture le sens de la phrase.
- **vecteur contextuel** : vecteur de taille fixe qui résume la phrase source et est utilisé pour guider la génération de la phrase cible.
- **décodeur** : génère la phrase cible à partir de cette représentation vectorielle, en produisant un mot à la fois.



Source : Sean Robertson, documentation PyTorch

Traduction automatique neuronale (NMT) : architecture encodeur-décodeur

Problème : utiliser un seul vecteur contextuel pour représenter une phrase source fonctionne avec les phrases courtes, mais devient problématique pour les phrases longues, car trop d'informations sont compressées dans un seul vecteur.

Traduction automatique neuronale (NMT) : architecture encodeur-décodeur

Problème : utiliser un seul vecteur contextuel pour représenter une phrase source fonctionne avec les phrases courtes, mais devient problématique pour les phrases longues, car trop d'informations sont compressées dans un seul vecteur.

Introduction du concept d'**attention** par **Bahdanau et al. (2016)**.

Traduction automatique neuronale (NMT) : attention

Introduction du concept d'**attention** par **Bahdanau et al. (2016)**.

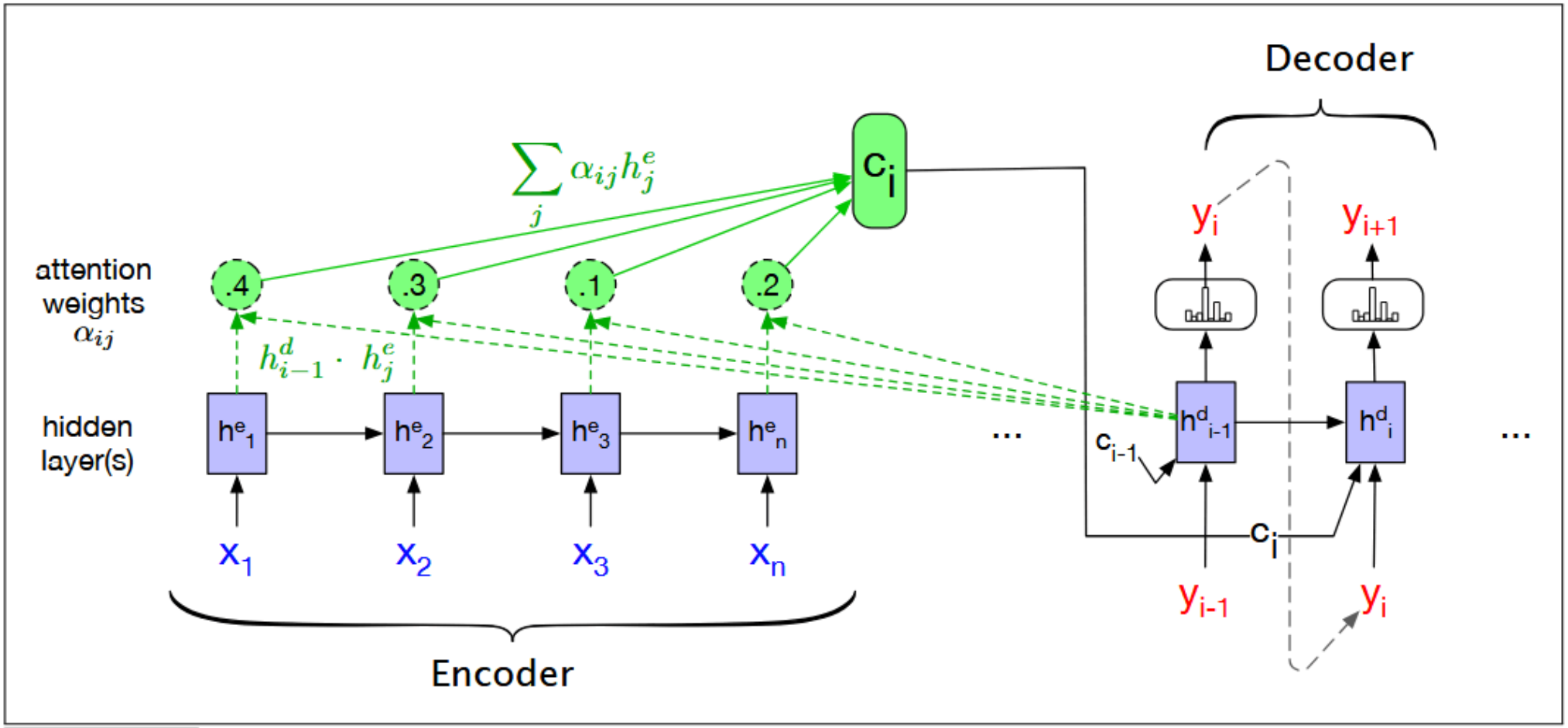
Au lieu d'un seul vecteur contextuel c, on utilise un ensemble de vecteurs contextuels c_t qui représentent les différentes parties de la phrase source à chaque étape t.

Traduction automatique neuronale (NMT) : attention

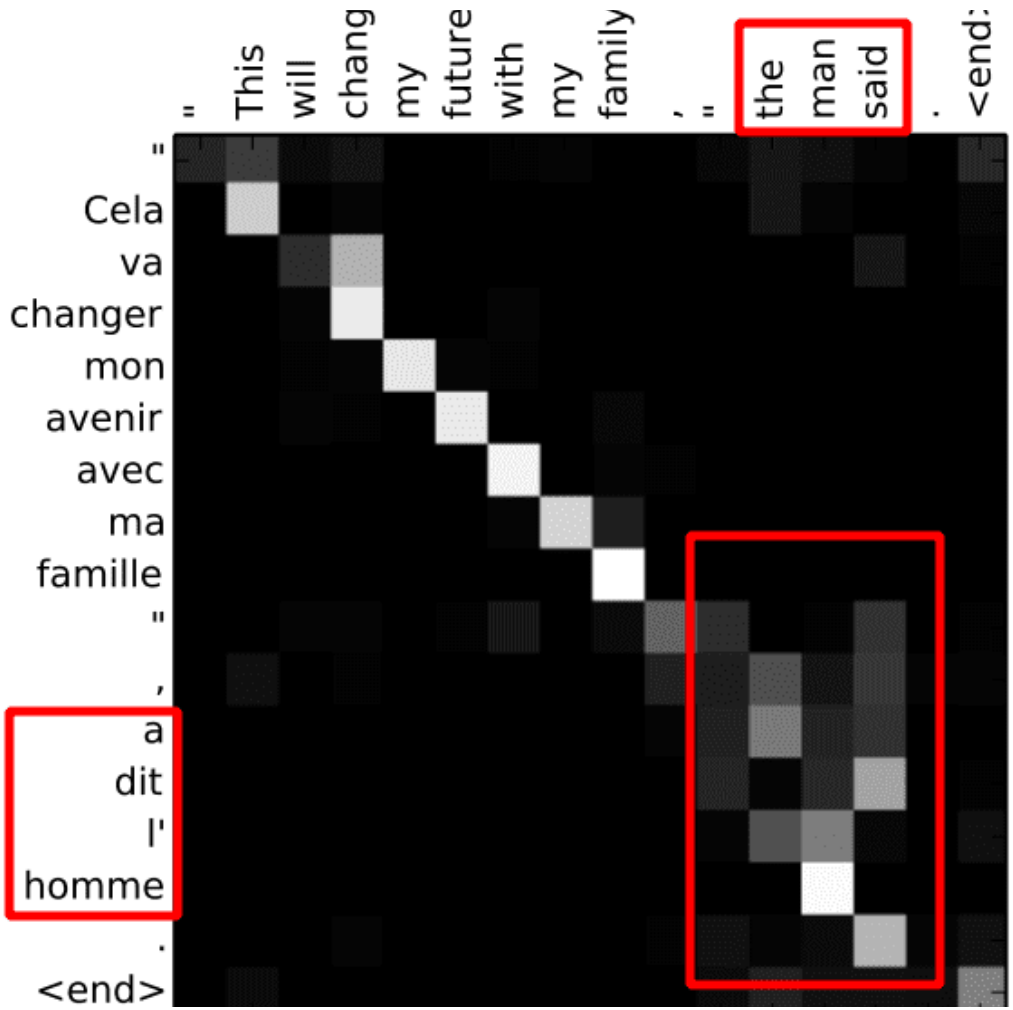
À chaque étape t de génération, le décodeur calcule un vecteur contextuel c_t comme combinaison pondérée des couches cachées :

$$c_t = \sum_j \alpha_{\text{ij}} h_j^e$$

Les coefficients d'attention α_{ij} sont calculés à partir des sorties de l'encodeur h_j^e et de la sortie précédente du décodeur h_i^d. Ils sont appelés **poids d'attention** et indiquent l'importance de chaque mot de la phrase source pour la génération du mot cible à l'étape t.



Source : **Jurafsky et Martin (2025)**

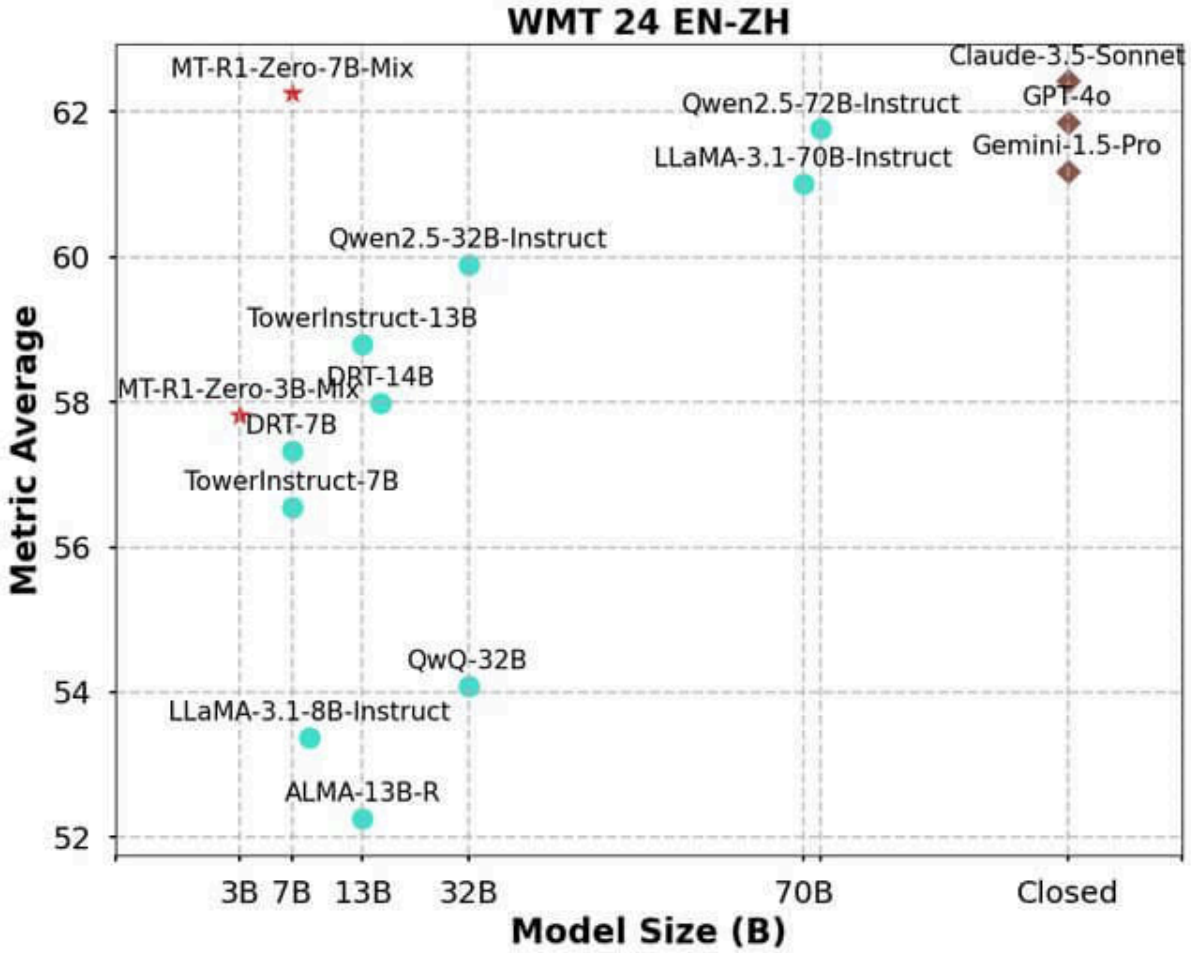


Exemple de changement d'ordre des mots

VI. La traduction automatique aujourd'hui

La traduction automatique aujourd'hui : les grands modèles de langue

- Recours de plus en plus important aux grands modèles de langue (LLM), notamment les LLM autorégressifs.
- Bonnes performances en *zero-shot* et *few-shot* learning, c'est-à-dire sans ou avec peu d'exemples de traduction.
- **Capacité multilingue** des modèles de langue, qui peuvent traduire entre plusieurs langues sans avoir été spécifiquement entrainés pour chaque paire de langues.
- Possibilité d'**affiner** les modèles sur ses propres données de traduction tout en conservant les capacités linguistiques du modèle entrainé ; essentiel pour la traduction spécialisée.


WMT 24 EN-ZH
Metric Average
62
60
58
56
54
52
MT-R1-Zero-7B-Mix
Claude-3.5-Sonnet
GPT-4o
Gemini-1.5-Pro
Qwen2.5-72B-Instruct
LLaMA-3.1-70B-Instruct
Qwen2.5-32B-Instruct
TowerInstruct-13B
DRT-14B
MT-R1-Zero-3B-Mix
DRT-7B
TowerInstruct-7B
QwQ-32B
LLaMA-3.1-8B-Instruct
ALMA-13B-R
3B
7B
13B
32B
70B
Closed
Model Size (B)

Modèles de langue

Nouvelles tâches de TA en lien avec les grands modèles de langue :

- stratégies de développement de prompts pour améliorer les traductions **(Pourkamali et Sharifi, 2024)** ;
- intégration de ressources lexicales dans les prompts **(Kim et al., 2024)** ;
- évaluation de la TA basée sur des LLM **(Qian et al., 2024)**.

Modèles de langue : avantages et limites

Avantages

- Possibilité d'utiliser les modèles sans données prélables.
- Possibilité de traduire des textes longs.
- Possibilité de définir un style de traduction ou des contraintes par du *prompt engineering*.

Limites

- Cout matériel, environnemental.
- Hallucinations ou non-respect des consignes.

Défis actuels de la TA

- Traduction depuis et vers des langues peu représentées dans les données d'entrainement (langues peu dotées ; **Song et al. (2025)**).

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	**Dahalo, Warlpiri, Popoloca, Wallisian, Bora**	2191	1.0 B	88.17%
	The Left-Behinds - Exceptionally limited resources, it will be a monumental, probably impossible effort to lift them up in the digital space.			
1	**Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo**	222	1.0 B	8.93%
	The Scraping Bys - Will take a solid, organized movement that increases awareness about these languages, and also sparks a strong effort to collect labeled datasets for them.			
2	**Zulu, Konkani, Lao, Maltese, Irish**	19	300 M	0.76%
	The Hopefuls - A small set of labeled datasets has been collected for these languages. Promising NLP tools can be created for these languages a few years down the line.			
3	**Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew**	28	1.1 B	1.13%
	The Rising Stars – With a strong web presence, there is a thriving cultural community online for them. However, they have been let down by insufficient efforts in labeled data collection.			
4	**Russian, Hungarian, Vietnamese, Dutch, Korean**	18	1.6 B	0.72%
	The Underdogs - A large amount of unlabeled data, comparable to those possessed by the winners, and are only challenged by a lesser amount of labeled data.			
5	**English, Spanish, German, Japanese, French**	7	2.5 B	0.28%
	The Winners - A dominant online presence, there have been massive industrial and government investments in the development of resources and technologies for these languages.			

Défis actuels de la TA

- Traduction depuis et vers des langues peu représentées dans les données d'entrainement (langues peu dotées ; **Song et al. (2025)**).
- Traduction multimodale **(Khan et al., 2024)**.
- Biais dans les données d'entrainement (p. ex. biais de genre : **Savoldi et al. (2025) ; Vanmassenhove (2024)**).
- Néologismes (**Lerner et Yvon (2025) ; Zheng et al. (2024)**).

Pour une revue plus complète, voir **Pang et al. (2025)**.

Bibliographie

Bahdanau, D., Cho, K., et Bengio, Y. (mai 2016). *Neural Machine Translation by Jointly Learning to Align and Translate* (Numéro arXiv:1409.0473). arXiv. 10.48550/arXiv.1409.0473

Chandioux, J. (août 1976). METEO, an Operational System for the Translation of Public Weather Forecasts. In D. G. Hays et J. Mathias (éds.), *Foreign Broadcast Information Service Seminar on Machine Translation: Foreign Broadcast Information Service Seminar on Machine Translation*.

Grass, T. (janvier 2010). À quoi sert encore la traduction automatique ?. *Cahiers du plurilinguisme européen*, *2*.

Hutchins, W. J. (2004). The Georgetown-IBM Experiment Demonstrated in January 1954. In *Lecture Notes in Computer Science: Lecture Notes in Computer Science* (p. 102‑114). Springer Berlin Heidelberg. 10.1007/978-3-540-30194-3_12

Jurafsky, D., et Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd éd.).

Khan, S., Tarun, A., Faraz, A., Kamble, P., Dahiya, V., Pokala, P., Kulkarni, A., Khatri, C., Ravi, A., et Agarwal, S. (novembre 2024). Chitranuvad: Adapting Multi-lingual LLMs for Multimodal Translation. In B. Haddow, T. Kocmi, P. Koehn, et C. Monz (éds.), *Proceedings of the Ninth Conference on Machine Translation: Proceedings of the Ninth Conference on Machine Translation*. 10.18653/v1/2024.wmt-1.80

Kim, S., Sung, M., Lee, J., Lim, H., et Perez, J. F. G. (octobre 2024). *Efficient Terminology Integration for LLM-based Translation in Specialized Domains* (Numéro arXiv:2410.15690). arXiv. 10.48550/arXiv.2410.15690

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press. 10.1017/CBO9780511815829

Koehn, P. (2020). *Neural Machine Translation*. Cambridge University Press.

Lerner, P., et Yvon, F. (janvier 2025). Towards the Machine Translation of Scientific Neologisms. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, et S. Schockaert (éds.), *Proceedings of the 31st International Conference on Computational Linguistics: Proceedings of the 31st International Conference on Computational Linguistics*.

Li, B. (novembre 2022). *Word Alignment in the Era of Deep Learning: A Tutorial*. **10.48550/arXiv.2212.00138**

Pang, J., Ye, F., Wong, D. F., Yu, D., Shi, S., Tu, Z., et Wang, L. (2025). Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models. *Transactions of the Association for Computational Linguistics*, *13*, 73‑95. **10.1162/tacl_a_00730**

Pinter, Y., Jacobs, C. L., et Bittker, M. (décembre 2020). NYTWIT: A Dataset of Novel Words in the New York Times. In D. Scott, N. Bel, et C. Zong (éds.), *Proceedings of the 28th International Conference on Computational Linguistics: Proceedings of the 28th International Conference on Computational Linguistics*. **10.18653/v1/2020.coling-main.572**

Pourkamali, N., et Sharifi, S. E. (janvier 2024). *Machine Translation with Large Language Models: Prompt Engineering for Persian, English, and Russian Directions* (Numéro arXiv:2401.08429). arXiv. **10.48550/arXiv.2401.08429**

Qian, S., Sindhujan, A., Kabra, M., Kanojia, D., Oŕasan, C., Ranasinghe, T., et Blain, F. (octobre 2024). *What Do Large Language Models Need for Machine Translation Evaluation?* (Numéro arXiv:2410.03278). arXiv. **10.48550/arXiv.2410.03278**

Savoldi, B., Bastings, J., Bentivogli, L., et Vanmassenhove, E. (juin 2025). A Decade of Gender Bias in Machine Translation. *Patterns*, *6*(6), 101257. **10.1016/j.patter.2025.101257**

Shen, L., Zhang, B., Matsoukas, S., Xu, J., et Weischedel, R. (octobre 2010). Statistical Machine Translation with a Factorized Grammar. In H. Li et L. Màrquez (éds.), *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Song, Y., Li, L., Lothritz, C., Ezzini, S., Sleem, L., Gentile, N., State, R., Bissyandé, T. F., et Klein, J. (juin 2025). *Is LLM the Silver Bullet to Low-Resource Languages Machine Translation?* (Numéro arXiv:2503.24102). arXiv. **10.48550/arXiv.2503.24102**

Steur, N., et Schwenker, F. (septembre 2021). Next-Generation Neural Networks: Capsule Networks with Routing-By-Agreement for Text Classification. *IEEE Access*, 1. **10.1109/ACCESS.2021.3110911**

Sumita, E., et Iida, H. (juin 1991). Experiments and Prospects of Example-Based Machine Translation. *29th Annual Meeting of the Association for Computational Linguistics*, 185‑192. **10.3115/981344.981368**

Vanmassenhove, E. (janvier 2024). *Gender Bias in Machine Translation and The Era of Large Language Models* (Numéro arXiv:2401.10016). arXiv. **10.48550/arXiv.2401.10016**

Weaver, W. (1949). Translation. In W. N. Locke et A. D. Boothe (éds.), *Machine Translation of Languages: Machine Translation of Languages* (p. 15–23). MIT Press.

Zheng, J., Ritter, A., et Xu, W. (août 2024). *NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms* (Numéro arXiv:2402.12261). arXiv. **10.48550/arXiv.2402.12261**

♡ Remerciements

- Pablo Ruiz Fabo pour le contenu de certaines diapositives.