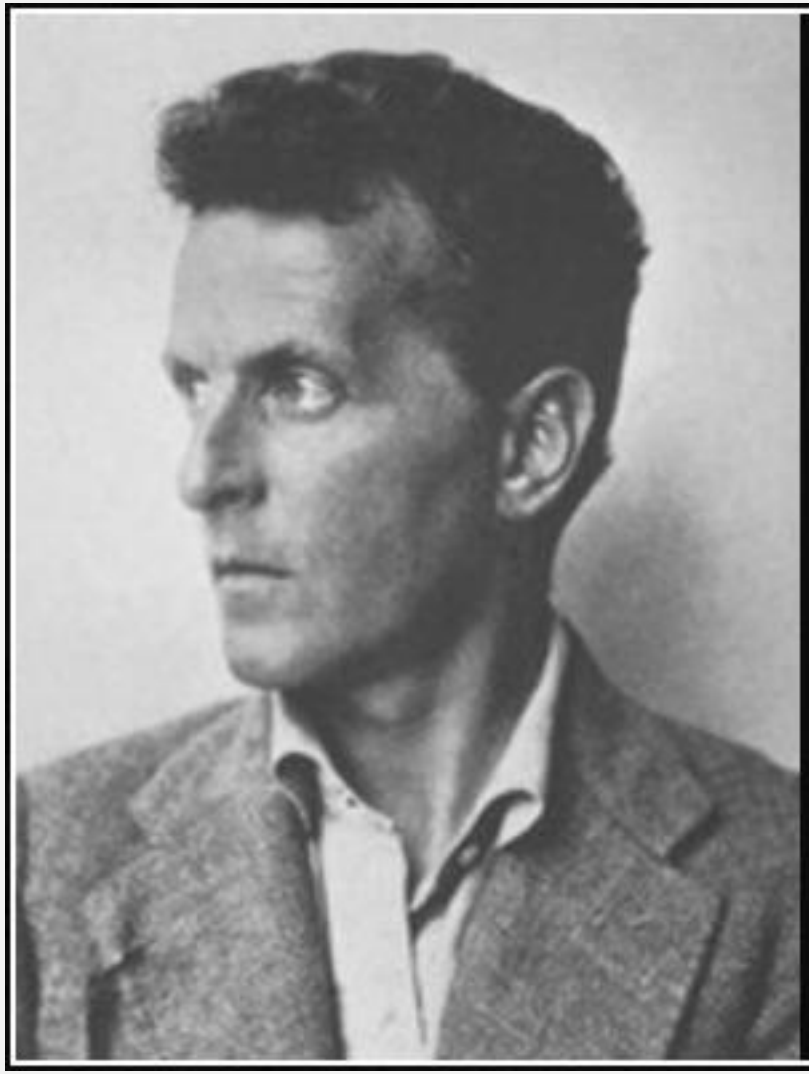
LA VECTORISATION SÉMANTIQUE ET L'HYPOTHÈSE DISTRIBUTIONNELLE

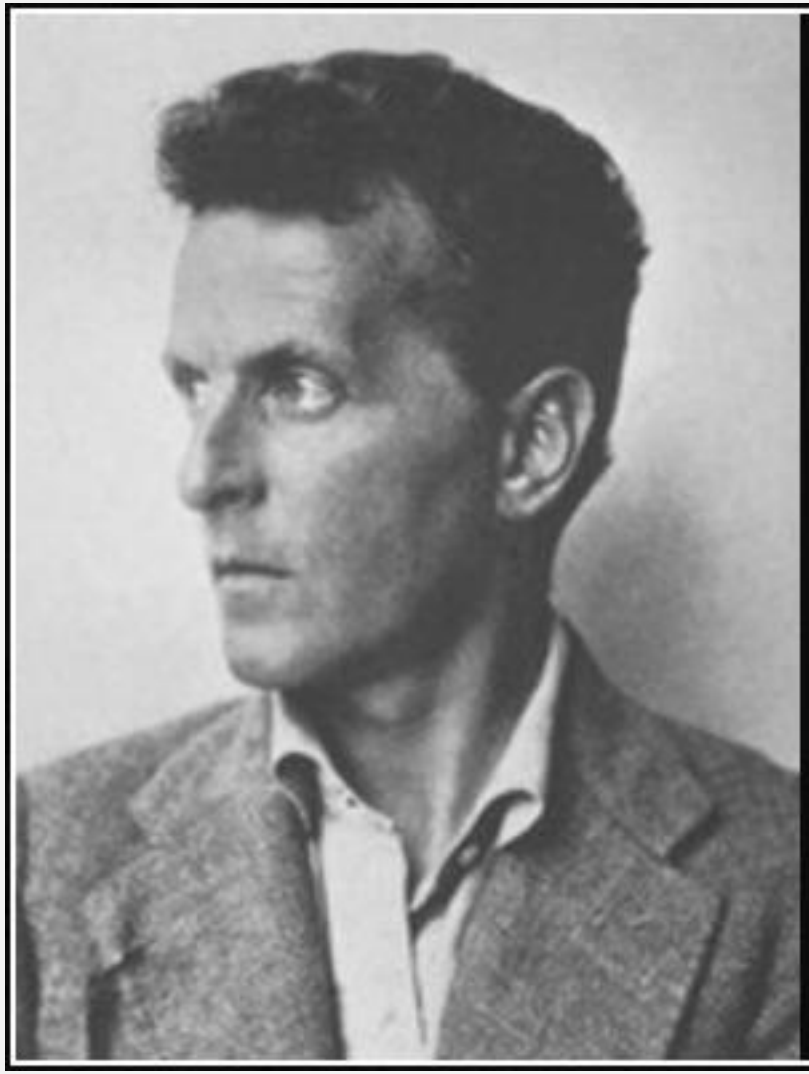
MEANING AS USE



For a large class of cases - though not for all - in which we employ the word meaning it can be explained thus: the meaning of a word is its use in the language.

— Ludwig Wittgenstein —

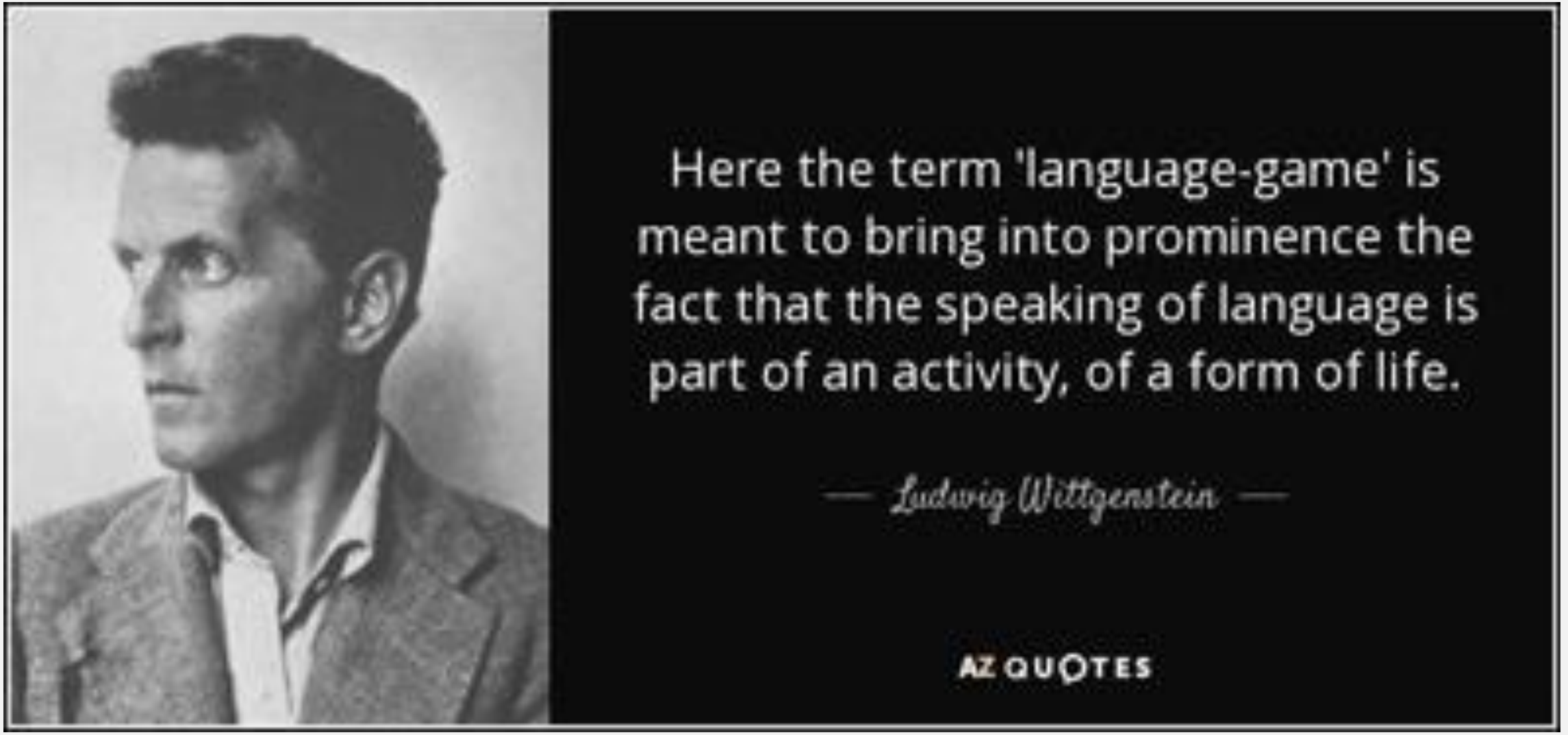
AZ QUOTES



The common behavior of mankind is the system of reference by means of which we interpret an unknown language.

— Ludwig Wittgenstein —

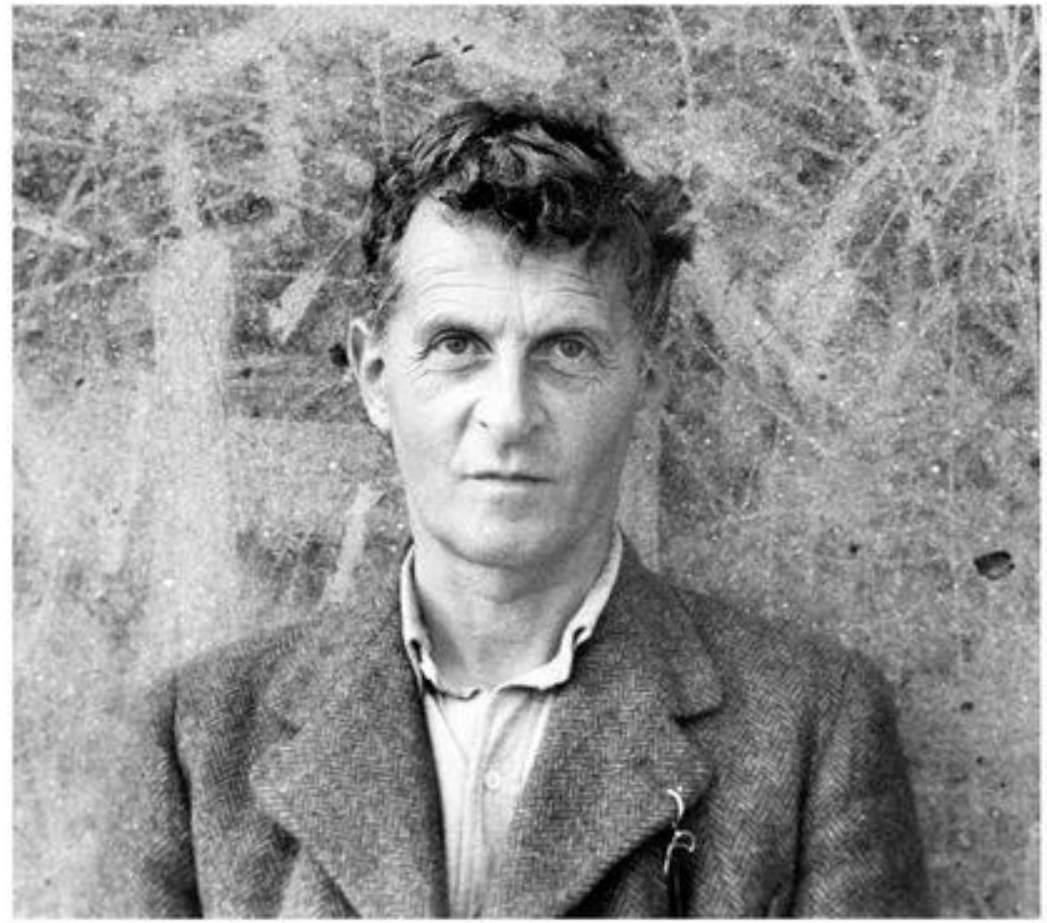
AZ QUOTES

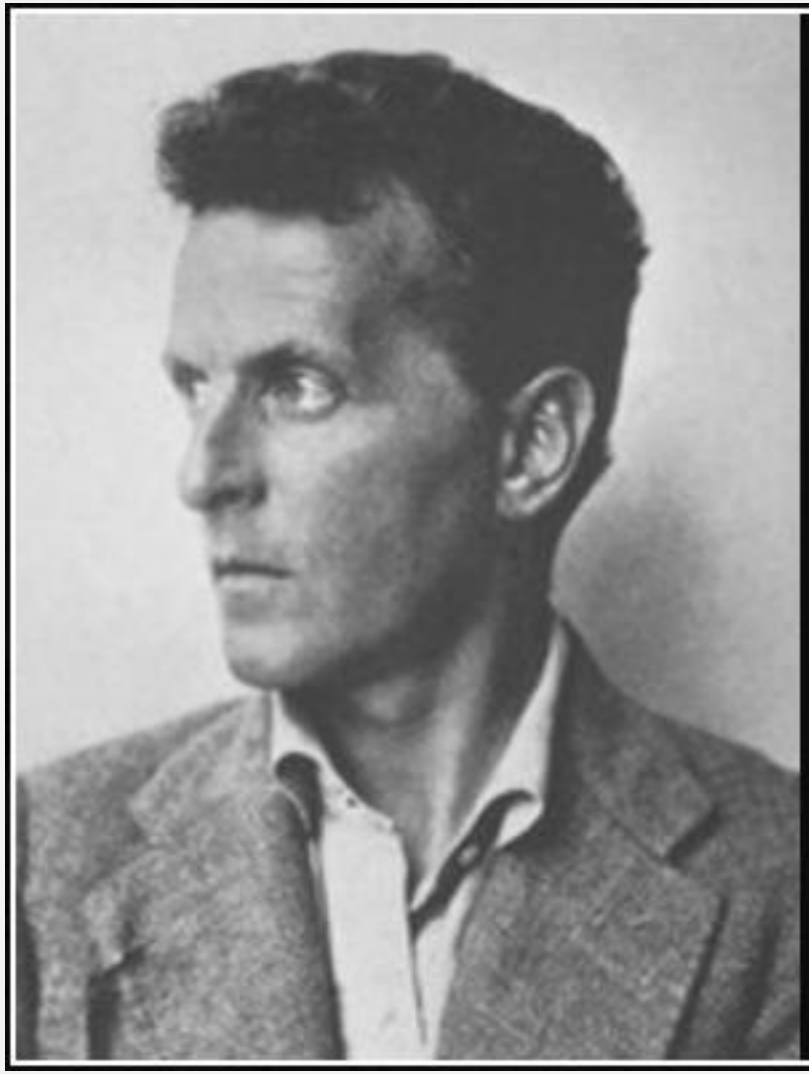

Here the term 'language-game' is meant to bring into prominence the fact that the speaking of language is part of an activity, of a form of life.
— Ludwig Wittgenstein —
AZ QUOTES

Wittgenstein vs. Representation

But what is the meaning of the word ‘five’?

--No such thing was in question here, only how the word ‘five’ is used.





When philosophers use a word--"knowledge," "being," "object," "I," "proposition," "name"--and try to grasp the essence of the thing, one must always ask oneself: is the word ever actually used in this way in the language-game which is its original home?--What we do is to bring words back from their metaphysical to their everyday use.

— Ludwig Wittgenstein —

AZ QUOTES

L'HYPOTHÈSE DISTRIBUTIONELLE

John Rupert Firth

"You shall know a word by the company it keeps"

-1957

- English linguist
- Most famous quote in NLP (probably)
- Modern interpretation: Co-occurrence is a good indicator of meaning

ZELLIG HARRIS

A THEORY OF LANGUAGE AND INFORMATION

A Mathematical Approach

CLARENDON PRESS
OXFORD

Studies in
NATURAL
LANGUAGE
PROCESSING

DISTRIBUTIONAL SEMANTICS

Alessandro Lenci and Magnus Sahlgren

Table 1
The most prominent vector models of semantic representation, along with short descriptions. More detailed descriptions are provided in later sections of the article.

Model	Authors	Year	Venue of Publication	Short Description
HAL	Lund & Burgess	1996	*Behavior Research Methods*	Creates a Word-by-Word Matrix
LSA	Landauer & Dumais	1997	*Psychological Review*	Creates a Word-by-Document Matrix and applies dimensionality reduction via SVD
Topic Models	Griffiths, Steyvers & Tenenbaum	2007	*Psychological Review*	Creates a Word-by-Document Matrix and applies dimensionality reduction via LDA
BEAGLE	Jones & Mewhort	2007	*Psychological Review*	Modifies initially random word vectors based on the other words in successively processed documents
word2vec	Mikolov, Chen, et al.; Mikolov, Sutskever, et al.	2013	*ICLR Workshop*, *Neural Information Processing Systems*	Trains word vectors as the hidden layer of a neural network that predicts words from the surrounding words, or vice versa
GloVe	Pennington, Socher & Manning	2014	*Empirical Methods in Natural Language Processing*	Trains word vectors to optimally predict words' probability of co-occurrence

VECTORISATION SEMANTIQUE

VECTORISATION DE MOTS (EMBEDDINGS)

- Le vectorisation d'un mot (ou d'un jeton, en fait un phonème) est un vecteur qui exprime, sous forme compressée, des informations sur les statistiques d'apparition de ce mot ou de ce jeton dans les données de manière plus générale

VECTORISATION DE MOTS
(EMBEDDINGS)

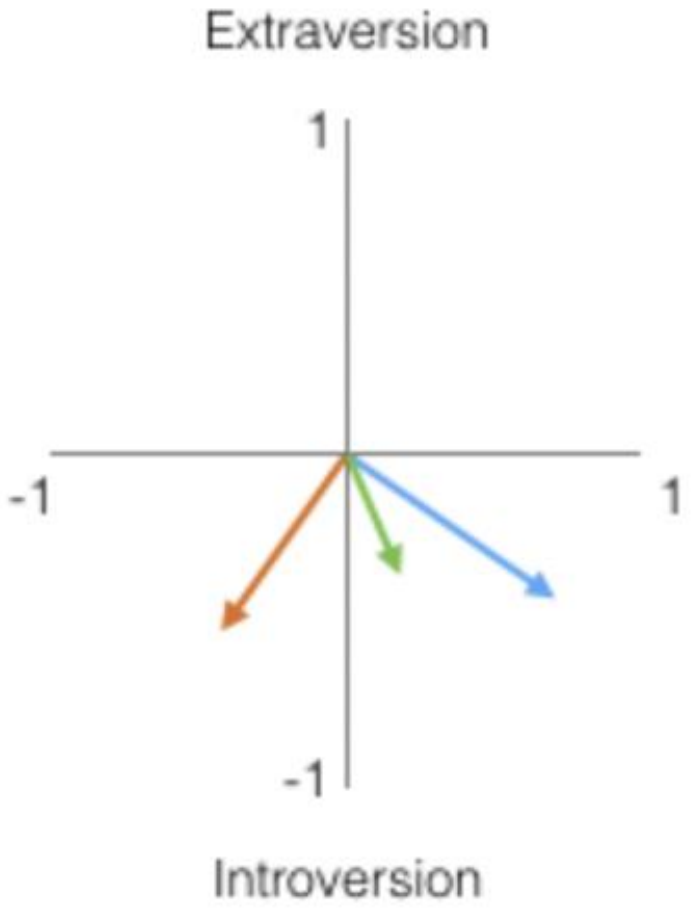
- Le résultat fascinant est qu'en utilisant cette méthode (en supposant que tu disposes d'un ensemble de textes suffisamment riche pour t'entraîner), tu arrives à des vecteurs qui capturent beaucoup de nos intuitions sémantiques sur les similitudes (sémantiques) entre les mots

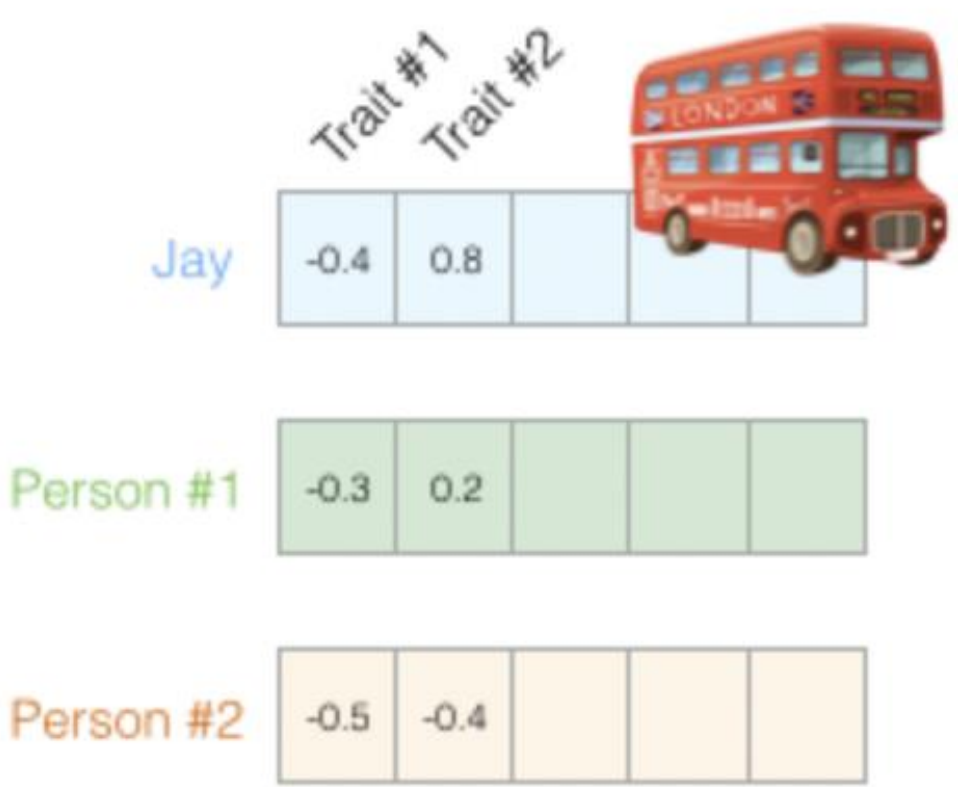
Personality Embeddings: What are you like?

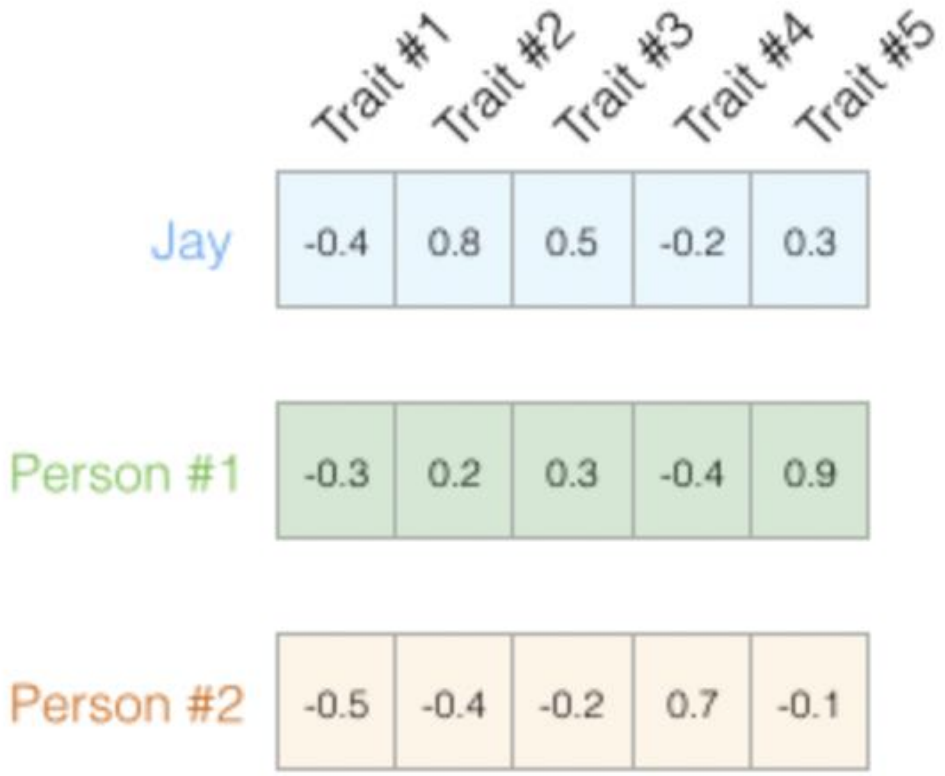
> *"I give you the desert chameleon, whose ability to blend itself into the background tells you all you need to know about the roots of ecology and the foundations of a personal identity" ~Children of Dune*

On a scale of 0 to 100, how introverted/extraverted are you (where 0 is the most introverted, and 100 is the most extraverted)? Have you ever taken a personality test like MBTI – or even better, the Big Five Personality Traits test? If you haven't, these are tests that ask you a list of questions, then score you on a number of axes, introversion/extraversion being one of them.

Openness to experience	79 out of 100
Agreeableness	75 out of 100
Conscientiousness	42 out of 100
Negative emotionality	50 out of 100
Extraversion	58 out of 100


Extraversion
1
-1
1
-1
Introversion

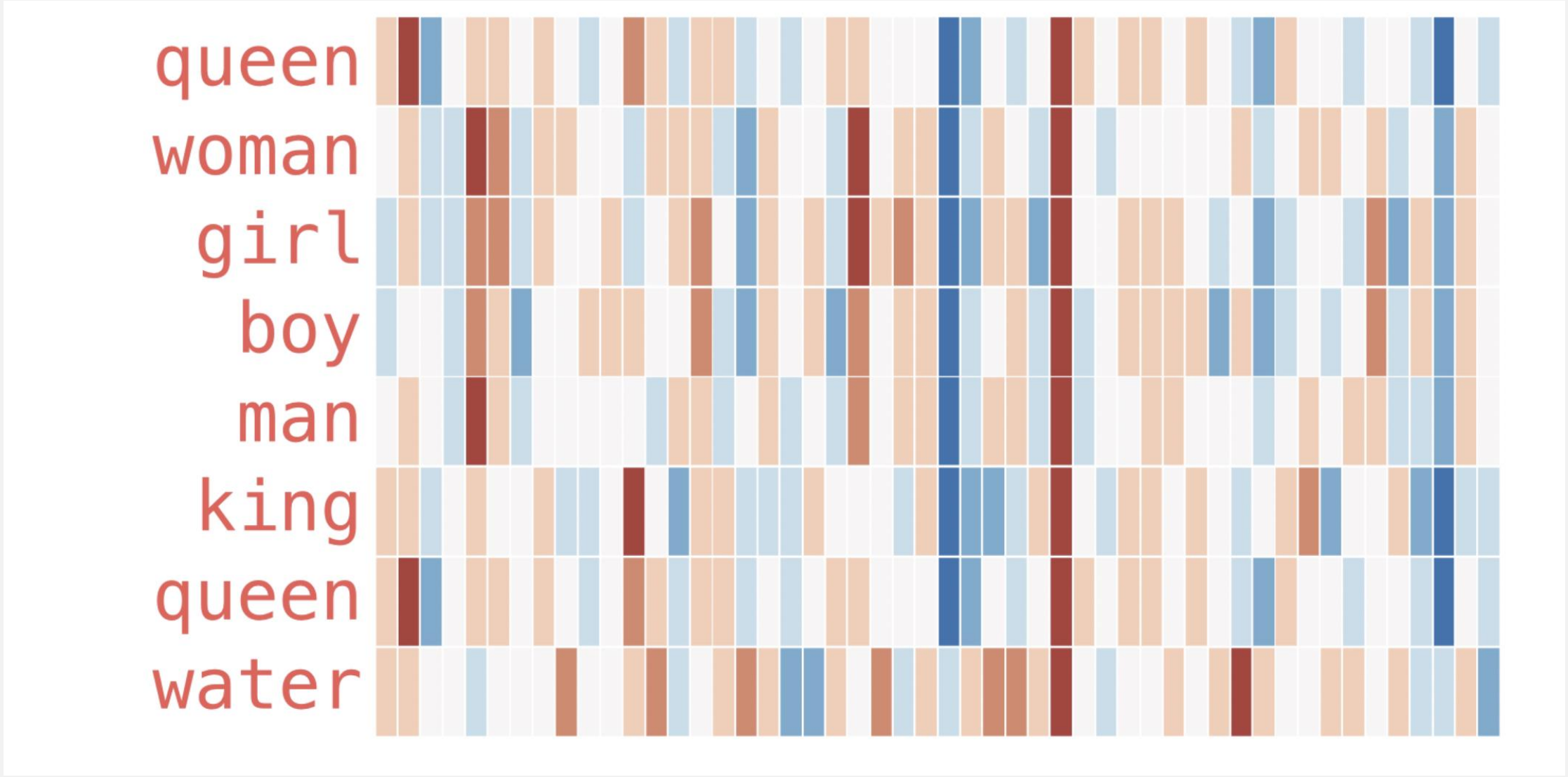

Trait #1
Trait #2
Jay
-0.4
0.8
Person #1
-0.3
0.2
Person #2
-0.5
-0.4


Trait #1
Trait #2
Trait #3
Trait #4
Trait #5
Jay
-0.4
0.8
0.5
-0.2
0.3
Person #1
-0.3
0.2
0.3
-0.4
0.9
Person #2
-0.5
-0.4
-0.2
0.7
-0.1

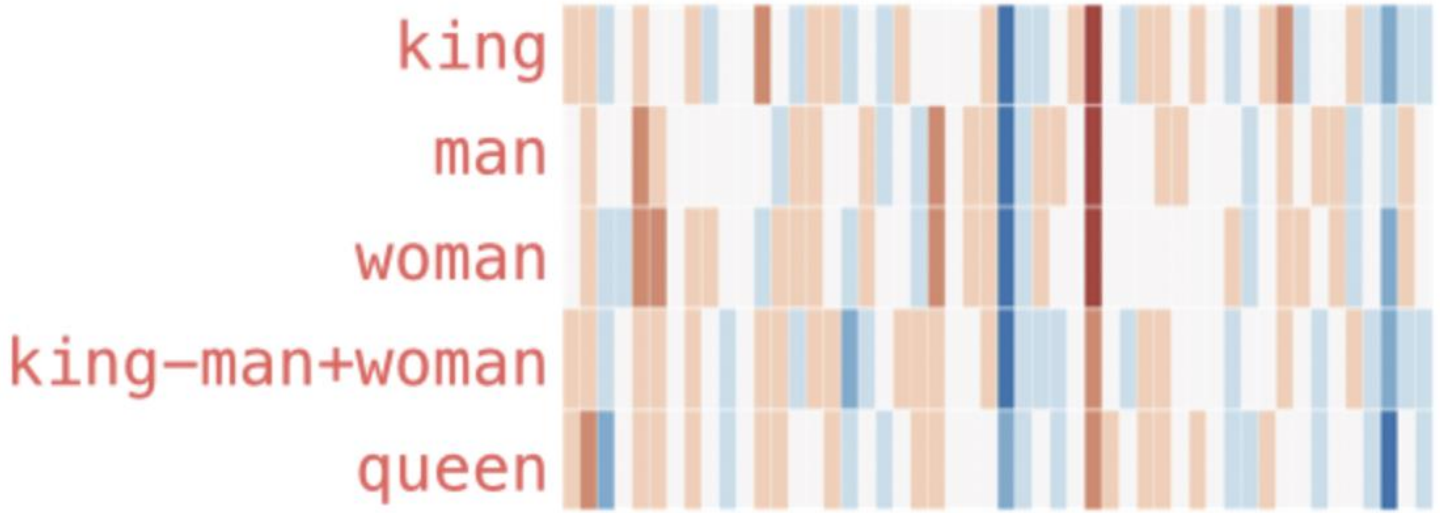
“king”

“Man”

“Woman”


queen
woman
girl
boy
man
king
queen
water

king - man + woman ~= queen



The resulting vector from "king-man+woman" doesn't exactly equal "queen", but "queen" is the closest word to it from the 400,000 word embeddings we have in this collection.

VECTORISATION

Mais comment un système d'intelligence artificielle peut-il apprendre de tels vecteurs ?

L'hypothèse de la distribution : la similarité sémantique est parallèle aux statistiques de cooccurrence des mots

VECTORISATION

Des mots sémantiquement similaires sont souvent utilisés ensemble ou à proximité l'un de l'autre

C'était une journée mouillée et pluvieuse

VECTORISATION

Étant donné que vous pouvez remplacer un synonyme par un autre, ils cooccurrent au même degré ou à un degré similaire avec un autre mot donné.

Le repas était délicieux

Le repas était succulent

VECTORISATION

C'est ainsi que les modèles d'IA qui traitent le langage naturel (RNN et Transformers) représentent les mots qu'ils traitent !

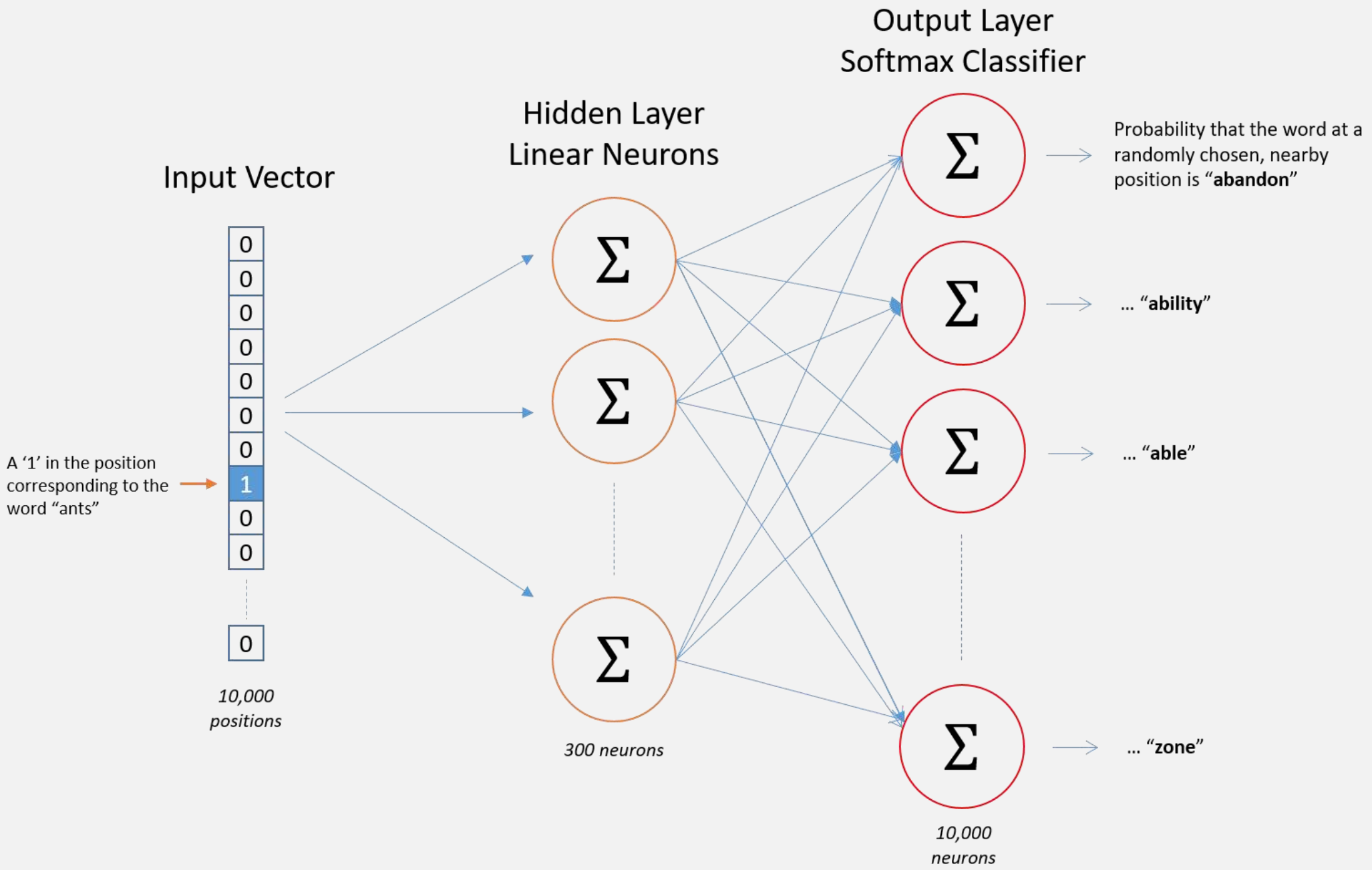
VECTORISATION

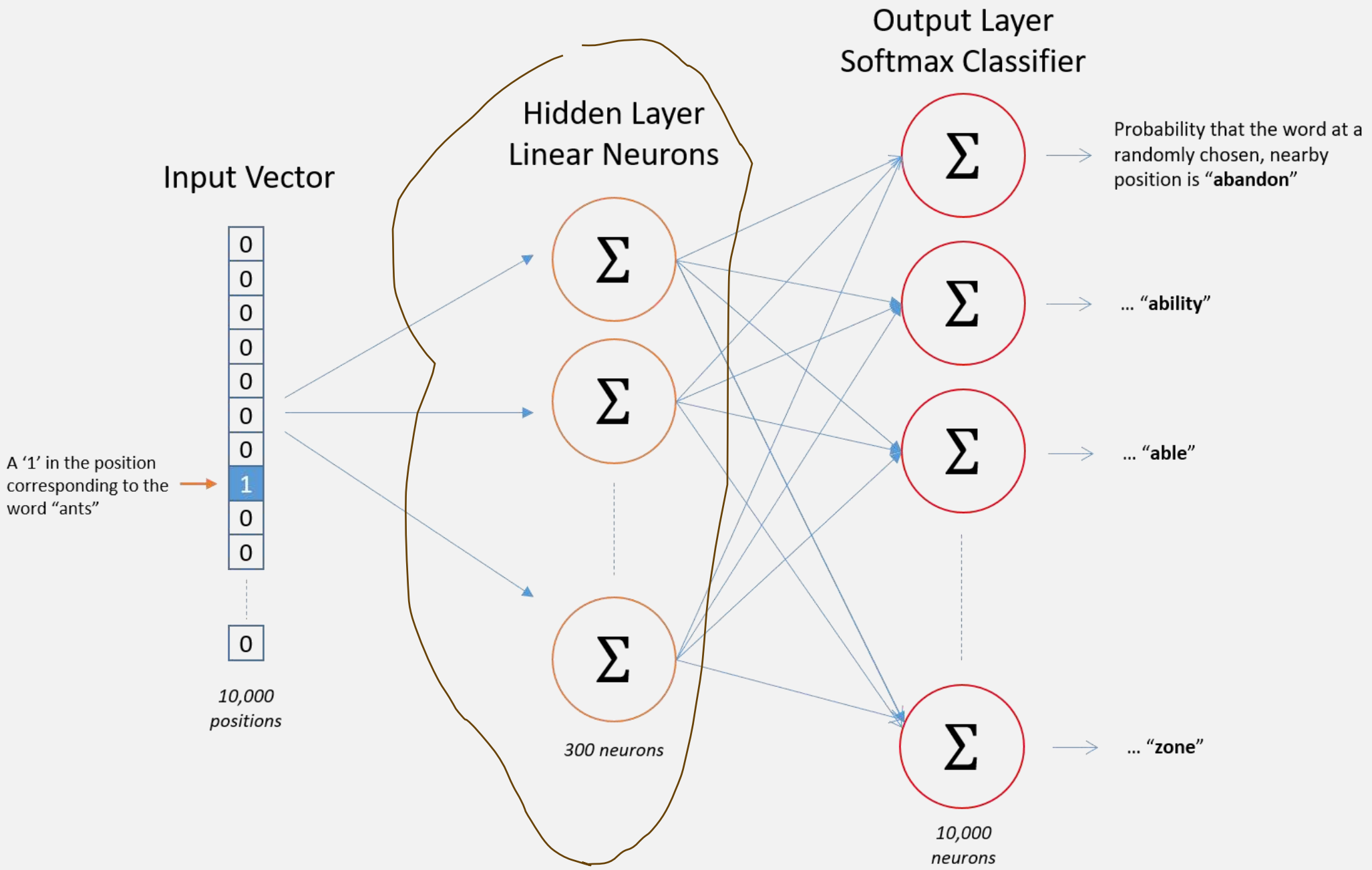
Il existe un type distinct d'algorithme d'apprentissage (pas forcement très profond), basée sur les travaux de Bengio, appelé modèle à Vectorisation (embedding model).

Parmi les exemples, citons *word2vec* et *ada* / *ada2* (ces derniers sont propriétaires d'OpenAI).

VECTORISATION

Il s'agit en fait d'algorithmes de compression.
Remarque : vous n'utilisez pas ces modèles pour leur sortie, vous obtenez les *vectorisations* de mots souhaités à partir de leur couche cachée.


Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
300 neurons
10,000 neurons


Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
300 neurons
10,000 neurons
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"

VECTORISATION

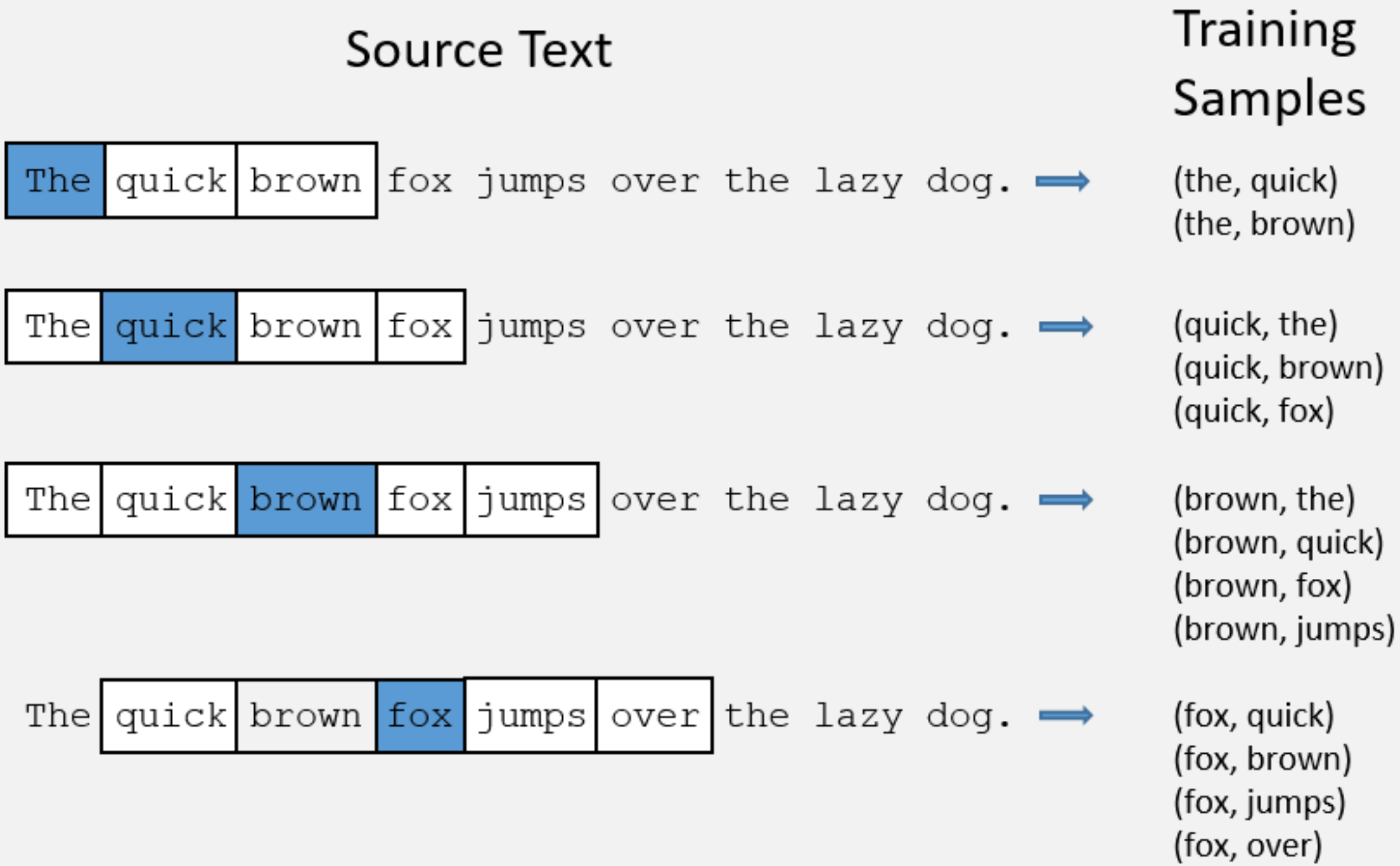
Entrée : un mot (représenté comme un vecteur "one-hot", essentiellement un index), [0,0,0,0,1,0,0...] si c'est le 5ieme mot

Sortie: une probabilité, pour tous les autres mots de l'index, de l'occurrence d'un mot à proximité.

VECTORISATION

Données d'entraînement pour la fonction de coût :

« contexts »: séquences de texte courtes, par exemple de 2 à 5 mots


Source Text
Training Samples
The quick brown fox jumps over the lazy dog.
(the, quick)
(the, brown)
The quick brown fox jumps over the lazy dog.
(quick, the)
(quick, brown)
(quick, fox)
The quick brown fox jumps over the lazy dog.
(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)
The quick brown fox jumps over the lazy dog.
(fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

VECTORISATION

Ce qui est étonnant, c'est que le modèle obtenu (même s'il ne comporte qu'une seule couche cachée) intègre de riches connaissances sémantiques !

```
king:queen::man:[woman, Attempted abduction, teenager, girl]
//Weird, but you can kind of see it

China:Taiwan::Russia:[Ukraine, Moscow, Moldova, Armenia]
//Two large countries and their small, estranged neighbors

house:roof::castle:[dome, bell_tower, spire, crenellations, turrets]

knee:leg::elbow:[forearm, arm, ulna_bone]

New York Times:Sulzberger::Fox:[Murdoch, Chernin, Bancroft, Ailes]
//The Sulzberger-Ochs family owns and runs the NYT.
//The Murdoch family owns News Corp., which owns Fox News.
//Peter Chernin was News Corp.'s COO for 13 yrs.
//Roger Ailes is president of Fox News.
//The Bancroft family sold the Wall St. Journal to News Corp.

love:indifference::fear:[apathy, callousness, timidity, helplessness, inaction]
//the poetry of this single array is simply amazing...
```

SENS, COMPRÉHENSION ET RÉFÉRENCE

- Logique et généralisation compositionnelle
- Contexte
- Ancrage sensorial / Incarnation

LOGIQUE ET GÉNÉRALISATION COMPOSITIONNELLE

- Notez que l'hypothèse wittgensteinienne de l'usage n'implique pas de sémantique distributionnelle
- Sémantique des rôles inférentiels : les significations des termes sont données par les inférences (déductives) que l'on peut en faire (Brandom).

CONTEXTE

- En général, ces méthodes ne tiennent pas compte du contexte, par exemple « apple » peut désigner le fruit ou la société informatique.

ANCRAGE SENSORIEL

- Chalmers considère l'argument suivant:
- 1. Les modèles linguistiques manquent de capacités sensorielles.
- 2. La pensée authentique nécessite des capacités sensorielles.
- Donc : 3. Les modèles linguistiques manquent de pensée authentique.

ANCRAGE SENSORIEL

- Thèse Sens-Pensée : Penser suppose d'avoir eu la capacité de sentir.

ANCRAGE SENSORIEL

- Qu'est-ce que la sensation ?
- Prendre des données ?
- Expérimenter ?
- Représenter ?

ANCRAGE SENSORIEL

- Chalmers :
- (1) Si les penseurs purs sont possibles, la pensée authentique ne nécessite pas de capacités sensorielles.
- (2) Les penseurs purs sont possibles.
- Ainsi, (3) la pensée authentique ne nécessite pas de capacités sensorielles.

ANCRAGE SENSORIEL

- Chalmers :
- (1) Si les penseurs purs sont possibles, la pensée authentique ne nécessite pas de capacités sensorielles.

- *Par definition, un penseur pur est quelqu'un capable de la pensée sans capacities sensorielles*

ANCRAGE SENSORIEL

- (2) Les penseurs purs sont possibles.

- *Les experiences de pensée de Avicenne et Descartes*

« Il faut que l'un de nous s'imagine qu'il a été créé d'un seul coup, et qu'il a été créé parfait, mais que sa vue a été voilée et privée de contempler les choses extérieures.

Qu'il a été créé tombant dans l'air ou dans le vide, de telle sorte que la densité de l'air ne le heurte, dans cette chute, d'aucun choc qui lui fasse sentir ou distinguer ses différents membres lesquels, par conséquent, ne se rencontrent pas et ne se touchent pas.

Eh bien ! qu'il réfléchisse et se demande s'il affirmera qu'il existe bien, et s'il ne doutera pas de son affirmation, de ce que son ipséité [c'est-à-dire son identité particulière] existe, sans affirmer avec cela une extrémité à ses membres, ni une réalité intérieure de ses entrailles, ni cœur, ni cerveau, ni rien d'entre les choses extérieures.

Bien mieux, il affirmera l'existence de son ipséité, mais sans affirmer d'elle aucune longueur, largeur ou profondeur.

Et s'il lui était possible, en cet état, d'imaginer une main ou un autre membre, il ne l'imaginerait ni comme une partie de son ipséité, ni comme une condition de son ipséité. Or tu sais bien, toi, que ce qui est affirmé est autre que ce qui n'est pas affirmé. Et la proximité est autre que ce qui n'est pas proche.

Par conséquent, cette ipséité dont est affirmée l'existence a quelque chose qui lui revient en propre, en ceci qu'elle est lui-même, par soi-même, non pas son corps et ses organes qui, eux, ne sont nullement affirmés.

Ainsi a-t-on l'occasion d'attirer l'attention sur une voie qui conduit à mettre en lumière l'existence de l'âme comme quelque chose qui est autre que le corps, mieux qui est autre que tout corps. Et que lui, il le sait et le perçoit.

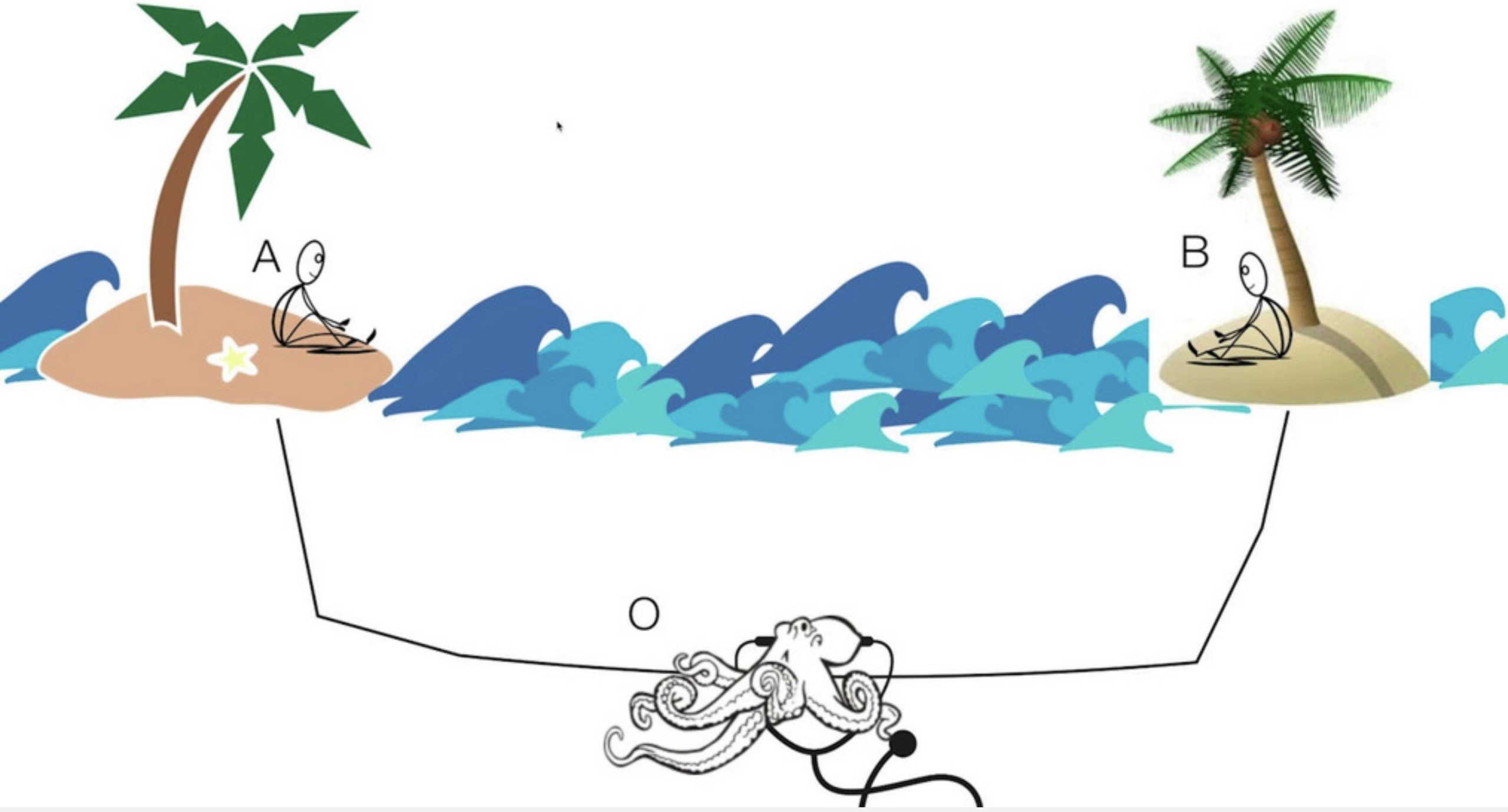
ANCRAGE SENSORIEL

- (2) Les penseurs purs sont possibles.

- *(a) Si l'homme flottante est concevable il est possible*
- *(b) il est concevable*
- *(c) donc, il est concevable*

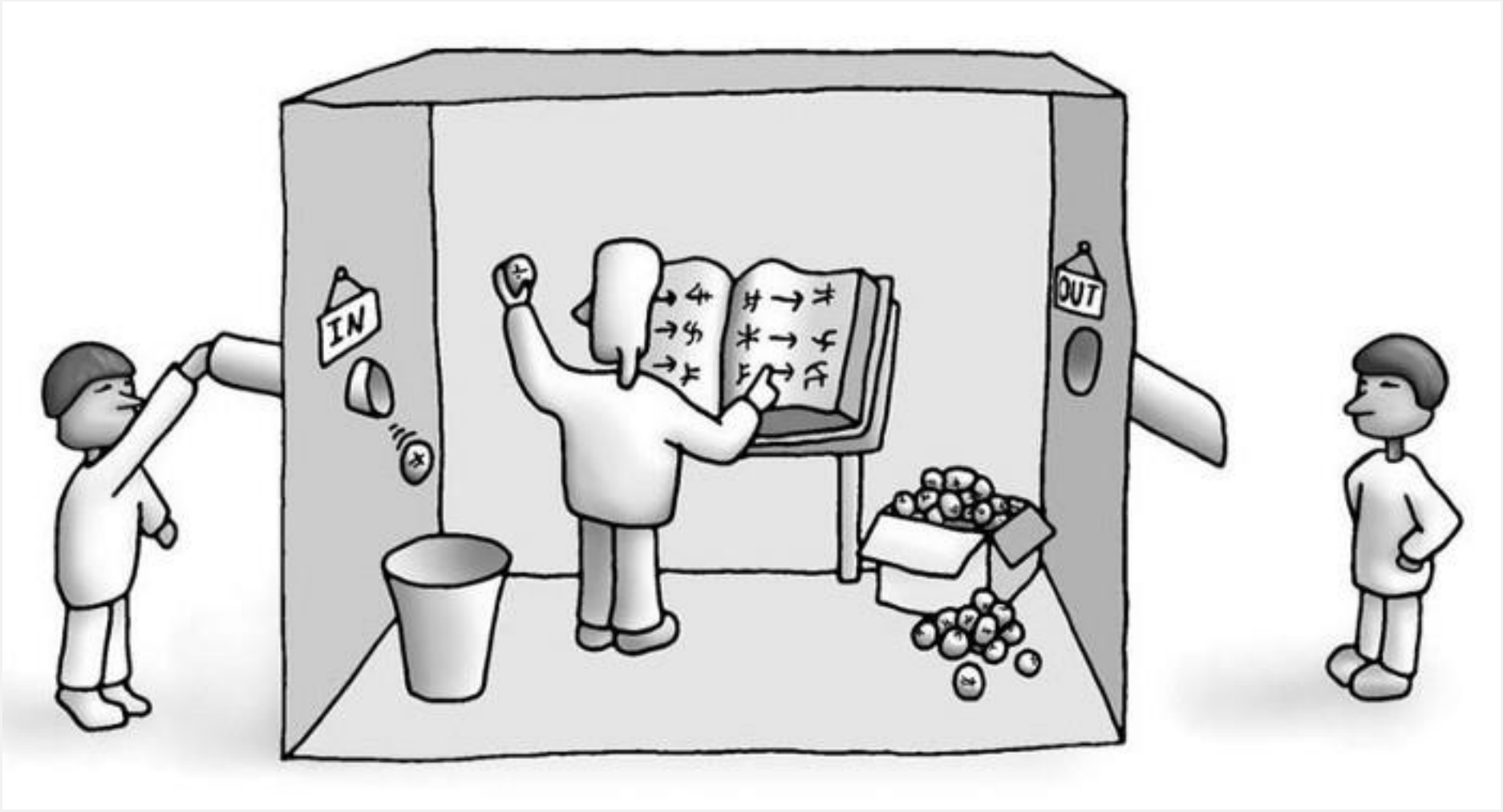
ANCRAGE SENSORIEL

- Bender et Koller 2020:


A
B
O

ANCRAGE SENSORIEL

- À comparer avec Searle (1980):


IN
OUT