- If A and B have almost identical environments we say that they are synonyms. *(Harris, 1954)*
- You shall know a word by the company it keeps. *(Firth, 1957)*
- The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear. *(Lenci, 2008)*

	QuatreVT 119 Kw	Voyage Bal 82 kw	Bête Hum. 128 kw	Mme Bovary 117 kw
bataille	35	4	6	2
clair	105	26	96	52
facile	12	19	6	10
politique	11	0	9	5
voyage	17	196	94	44
idiot	2	1	2	6
amour	19	0	47	94

TABLE K.1 – Matrice terme-documents pour quelques mots et 4 romans (Quatrevingt-treize (Hugo) ; Le voyage en ballon (Verne) ; La bête humaine (Zola) ; Mme Bovary (Flaubert)).

QuatreVT	(17,19)
Voyage Bal	(196,0)
Bête Hum.	(94,47)
Mme Bovary	(44,94)

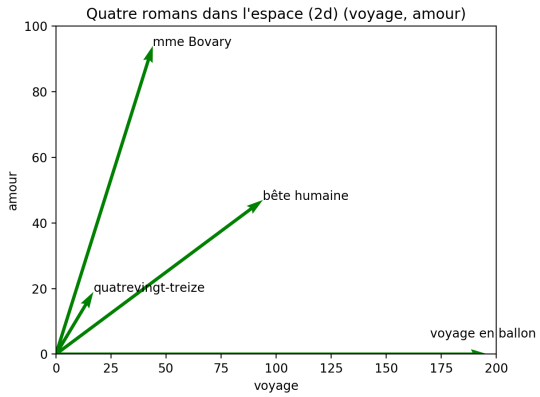

FIGURE K.9 – Représentation graphique de quelques romans dans le plan (voyage, amour)

amour	(19,94)
bataille	(35,2)
facile	(12,10)
politique	(11,5)
voyage	(17,44)

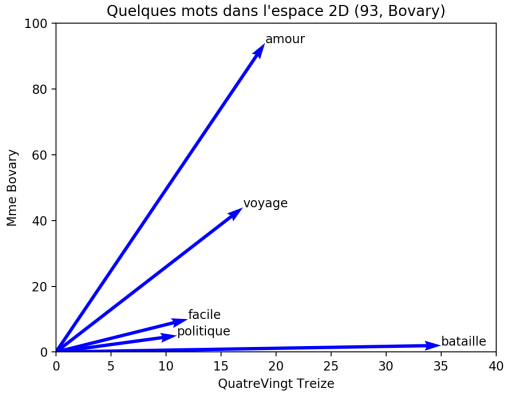

FIGURE K.10 – Représentation graphique de quelques mots dans le plan (93, Bovary)

Pondération	*Formule du tf*
binaire	$\begin{cases} 1 & \text{si } f_{t,d} > 0 \\ 0 & \text{sinon} \end{cases}$
fréquence brute	$f_{t,d}$
fréquence (normalisée)	$\frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$
normalisation logarithmique	$\log(1 + f_{t,d})$
normalisation par le max	$K + (1-K)\frac{f_{t,d}}{max_{t' \in d}\, f_{t',d}}$

Formule pour idf
$idf_t = \log\left(\frac{\lvert D \rvert}{df_t}\right)$

Formule complète
$td\text{-}idf_{t,d} = tf_{t,d} \times idf_t$

TABLE K.2 – Formules pour le calcul du tf-idf, avec les variantes les plus courantes pour le calcul du tf.
Notations : t terme
D ensemble de documents, $d \in D$
$f_{t,d}$ fréquence de t dans d
df_t $|\{d \in D \ / \ f_{t,d} \neq 0\}|$

woman − queen + man		queen − woman + king		king − man + queen		man − king + woman	
king	0.749	man	0.733	woman	0.745	son	0.728
prince	0.708	who	0.712	beautiful	0.726	queen	0.716
kingdom	0.694	men	0.711	girl	0.672	elizabeth	0.710
victoria	0.644	whom	0.708	my	0.655	brother	0.706
scotland	0.643	killed	0.685	lady	0.646	emperor	0.699
wales	0.640	person	0.676	she	0.628	wife	0.693
lord	0.638	young	0.673	thing	0.613	henry	0.691
great	0.627	himself	0.639	good	0.606	younger	0.686
elizabeth	0.622	said	0.637	her	0.595	daughter	0.681
throne	0.614	father	0.636	naked	0.592	prince	0.665

TABLE K.3 – Quelques calculs analogiques réalisés avec la démo de GloVe (`https://github.com/stanfordnlp/GloVe`) avec les 10 mots les plus proches par distance cosinus du point obtenu par l'opération additive sur les vecteurs.

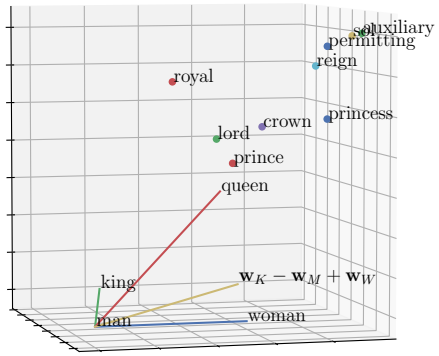

Figure 1. The relative locations of word embeddings for the analogy "*man* is to *king* as *woman* is to ..?". The closest embedding to the linear combination $\mathbf{w}_K - \mathbf{w}_M + \mathbf{w}_W$ is that of *queen*. We explain why this occurs and interpret the difference between them.

FIGURE K.11 – Graphique sur la fameuse analogie homme/femme|roi/x, version de Allen et Hospedales (2019)