Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Évaluer la capacité des transformeurs à distinguer les significations compositionnelles et idiomatiques d'une même expression

Nina Nusbaumer[1] Guillaume Wisniewski[1] Benoît Crabbé[1]
(1) Université Paris Cité, LLF, CNRS, 75013 Paris, France
nina.nusbaumer,guillaume.wisniewski,benoit.crabbe@u-paris.fr

RÉSUMÉ

Cet article explore comment les modèles de langue fondés sur les transformeurs encodent les significations compositionnelles et non-compositionnelles de séquences en anglais comme *big fish*, qui, selon le contexte, peuvent signifier soit « grand poisson », soit « personne importante ». Nous avons mené des expériences pour évaluer : (1) la distinction entre les plongements lexicaux des groupes nominaux compositionnels et non compositionnels à travers les couches du modèle de langue, (2) leur séparabilité linéaire, et (3) l'unité lexicale des séquences non compositionnelle. Nos résultats montrent que le modèle différencie bien les deux significations, et ce dès les premières couches, avec néanmoins une variabilité selon les expressions. De plus, s'appuyant sur des informations contextuelles plus larges, le modèle ne traite pas les expressions idiomatiques comme lexicalement plus unifiées que leurs équivalents compositionnels.

ABSTRACT

Assessing Transformers' Ability to Distinguish between the Compositional and Idiomatic Meanings of the Same Expression

This paper explores how transformer-based LMs encode compositional (C) and non-compositional (NC) meanings of sequences like *big fish*, which, depending on the context, can either mean "large fish" or "important person". We conducted experiments to evaluate : (1) the distinction between C and NC embeddings across LM layers, (2) their linear separability, and (3) the lexical unity of NC sequences. Our results demonstrate that the model differentiates C and NC meanings, with deeper layers showing improved differentiation. While linear separability is often achieved in early layers, variability exists across expressions. The LM does not however treat idiomatic expressions as lexically unified but relies on broader contextual information.

MOTS-CLÉS : Expressions polylexicales, compositionalité, idiomaticité, transformeurs, représentations sémantiques.

KEYWORDS: Multiword Expression (EP), compositionality, idiomaticity, transformers, semantic representations.

1 Introduction

Les transformeurs peuvent-ils distinguer les significations idiomatiques des significations non idiomatiques d'un même groupe de mots ? Pour répondre à cette question, nous examinons, dans ce travail,

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 360–375.

comment ces modèles représentent (en anglais) les syntagmes nominaux non compositionnels (dont le sens est idiomatiques) par rapport à leurs équivalents compositionnels (dont le sens est littéral et se déduit du sens des mots le composant), comme dans l'exemple ci-dessous :

(1) The kid catches a *big fish* in the pound.

L'enfant attrape un *grand poisson* dans l'étang.

(2) She's a *big fish* in the company.

C'est une *personne importante* dans l'entreprise.

Ici, un même syntagme nominal *(big fish)* possède deux significations distinctes. Dans le premier cas, son sens découle de manière compositionnelle des significations des deux mots individuels. Dans le second cas, il s'agit d'une expression idiomatique, dont le sens ne peut être déduit des mots qui la composent et doit être interprété de manière non-compositionnelle. Ce deuxième type de syntagme nominal appartient à la catégorie plus large des expressions polylexicales comme par exemple les idiomes (p.ex. « avoir un cœur d'artichaut »), les expressions binomiales (p.ex. « ici et maintenant »), ou encore les noms composés — cette dernière catégorie étant au centre de cette étude. Notre travail aborde un défi majeur en TAL : comprendre comment les transformeurs (Vaswani *et al.*, 2017) traitent la compositionnalité, une caractéristique fondamentale de la langue humaine (Szabó, 2020; Liu & Neubig, 2022), au travers de leurs représentations internes.

Nos expériences révèlent trois caractéristiques clés des modèles de type transformeur, en particulier `BERT` (Devlin *et al.*, 2019) : (1) ce modèle distingue systématiquement les significations compositionnelles et non-compositionnelles à travers ses couches ; (2) cette différenciation apparaît dès la première couche, suggérant un processus précoce de désambiguïsation ; (3) la capacité du modèle à discriminer ces significations varie selon les couches et les expressions polylexicales, ce qui reflète la nature continue et dépendante du contexte de la compositionnalité. Plus précisément, dans cet article, nous cherchons à déterminer si les transformeurs sont capables d'identifier les cas où le sens d'un composé nominal doit être construit de manière compositionnelle ou non, directement au niveau des représentations. L'étude de cette distinction est essentielle pour mieux comprendre le fonctionnement interne des modèles et leurs limites face aux phénomènes de compositionalité et d'idiomaticité.

Le reste de cet article est organisé de la manière suivante : la section 2 présente la problématique et les travaux connexes, la section 3 décrit les données, les expériences et les résultats, et la section 4 conclut en discutant les implications et les perspectives futures [1].

2 Contexte

Les expressions polylexicales fonctionnent, dans une certaine mesure, comme des unités de sens uniques (Baldwin & Kim, 2010; Constant *et al.*, 2017; Dankers *et al.*, 2022). À ce titre, elles posent un défi aux modèles de TALN, car : (1) elles peuvent être polysémiques, (2) elles couvrent un spectre

1. Le code ainsi que le corpus sont disponibles ici : https://github.com/NinaNusb/eval-transformers-idiomaticity.git.

allant de totalement compositionnelles (par exemple « changement climatique ») à hautement non compositionnelles (par exemple « cheville ouvrière »), et (3) elles apparaissent sous leur forme composée moins fréquemment que leurs composants pris individuellement. Alors que les humains traitent sans effort les expressions polylexicales non-compositionnelles comme des unités holistiques stockées en mémoire (Bhattasali *et al.*, 2019), les modèles computationnels doivent relever un défi complexe : formaliser une sémantique capturant à la fois les aspects compositionnels et non compositionnels.

Afin de représenter efficacement un large éventail de structures linguistiques tout en maintenant un vocabulaire réduit, les modèles transformeurs s'appuient sur des mécanismes qui encouragent une interprétation compositionnelle, tant au niveau des mots qu'à celui des sous-mots, en construisant le sens à partir de la combinaison systématique d'unités élémentaires (Tayyar Madabushi *et al.*, 2021; Devlin *et al.*, 2019). Cependant, cette architecture entre en conflit avec la nature des expressions polylexicales, en particulier les expressions idiomatiques (Hashempour & Villavicencio, 2020; Mitchell & Lapata, 2010), dont la sémantique ne peut être dérivée à partir des sens de leurs composants. Dès lors, les modèles pré-entraînés peinent à encoder de manière cohérente ces expressions, car leur conception même privilégie la construction dynamique des représentations sémantiques à partir des unités lexicales, plutôt que l'apprentissage de significations figées et non compositionnelles (Mitchell & Lapata, 2010; Mikolov *et al.*, 2013). Bien que ces modèles puissent, dans une certaine mesure, prédire la représentation d'une expression à partir de ses constituants, leurs performances ne correspondent pas aux jugements humains sur la compositionnalité sémantique (Liu & Neubig, 2022). La diversité des méthodes d'évaluation et des définitions employées dans la littérature a conduit à des conclusions divergentes sur la capacité des modèles à modéliser la compositionnalité, en l'absence d'un consensus sur ce qui caractérise un comportement compositionnel dans ces architectures (Hupkes, 2020).

Les recherches portant sur le traitement des expressions figées dans les transformeurs ont principalement exploré la distinction entre idiomes et paraphrases. Par exemple, Tian *et al.* (2023) montrent que la similarité inter-couches des représentations d'idiomes et de leurs reformulations évolue au fil des couches, suggérant un encodage dynamique des expressions idiomatiques. Toutefois, le recours aux paraphrases introduit un facteur confondant qui limite l'évaluation fine des distinctions sémantiques. Dans cette étude, nous proposons une approche complémentaire visant à isoler la capacité des modèles à encoder des divergences de sens à partir du seul contexte, sans appui sur des reformulations : nous examinons comment les transformeurs distinguent les sens littéraux et idiomatiques d'expressions nominales homonymes, dans un cadre entièrement non supervisé.

D'autres travaux ont souligné les limites des transformeurs dans la représentation des expressions figées. Zeng & Bhat (2022) montrent que `BART` (Lewis *et al.*, 2019) peine à apprendre des représentations idiomatiques, même après un entraînement spécifique sur ce type d'expressions. Néanmoins, leur approche repose sur un apprentissage supervisé, dans lequel le modèle est explicitement entraîné à reconstruire des expressions masquées. À l'inverse, notre objectif est d'analyser les représentations spontanément produites par un modèle pré-entraîné, tel que `BERT`, afin d'évaluer directement sa sensibilité contextuelle aux différentes lectures d'une même expression polylexicale, sans supervision ni *fine-tuning*.

Enfin, la fréquence des usages littéraux des expressions polylexicales est souvent considérée comme un facteur déterminant dans leur traitement par les modèles. Savary *et al.* (2019) observent que les occurrences littérales des expressions figées sont rares, ce qui soulève la question de leur pertinence dans l'analyse computationnelle de la langue. Toutefois, cette conclusion repose principalement sur l'étude des idiomes, où l'interprétation figurée est largement dominante. En revanche, les composés

nominaux présentent une plus grande variabilité d'usage : certains, comme *big fish* ou *gold mine* [2], sont fréquemment employés dans leurs sens littéral et idiomatique. Leur analyse constitue ainsi un cadre pertinent pour examiner la capacité des modèles à différencier des interprétations contextuelles distinctes.

3 Représentation des Expressions Polylexicales dans BERT

Nous présentons ici trois expériences conçues pour déterminer si les représentations des expressions polylexicales (EPs) dans les modèles de transformeurs varient selon que leur signification est construite compositionnellement ou non.

Jeu de données et modèle Nous utilisons la partie anglaise [3] du jeu de données `AStitchInLanguageModels` (Tayyar Madabushi *et al.*, 2021), qui fournit des exemples d'usages à la fois compositionnels et idiomatiques des mêmes EPs. Ce corpus se concentre sur les noms composés et contient 648 exemples avec trois phrases par exemple (une contenant l'EP cible et deux phrases de contexte). Le jeu de données complet comprend 4 645 exemples de 223 EPs, mais nous excluons les noms propres, les usages métaphoriques et les exemples n'ayant qu'un seul type de signification afin de permettre des comparaisons pertinentes. Au final, nous considérons, dans nos expériences, 38 EPs uniques associant un nom avec un adjectif (p.ex. *silver spoon* [4]), un verbe (p.ex. *closed book* [5]) ou un autre nom (p.ex. *brick wall* [6]). Chaque exemple est annoté comme compositionnel (C) ou non-compositionnel (NC), comme illustré ci-dessous [7] :

(3) (C) That doesn't mean, however, that we're not interested in your *big fish* photos.

« Cela ne veut pas dire, cependant, que nous ne sommes pas intéressés par vos photos de *grands poissons*. »

(4) (NC) The pending UFA list lacks *big fish* players on non-contending teams.

« La liste des UFA en attente manque de joueurs *importants* dans les équipes non compétitives. »

Nous utilisons `BERT` (Devlin *et al.*, 2019) pour encoder les trois phrases. [8]. La tokenisation de chaque EP en deux mots produit deux tokens, dont nous extrayons et concaténons les représentations, obtenant ainsi un vecteur de 1 536 dimensions pour chaque exemple.

2. Littéralement « mine d'or », pouvant aussi signifier « source extrêmement riche en ressources, en opportunités ou en informations précieuses ».
3. Il n'existe pas, à notre connaissance, un jeu de données similaire en français.
4. Littéralement « cuillère en argent », pouvant aussi signifier « privilège ».
5. Littéralement « livre fermé », pouvant aussi signifier « difficile à comprendre ».
6. Littéralement « mur de briques », pouvant aussi signifier « un problème ou une situation difficile à résoudre ».
7. Seule la phrase contenant l'EP est affichée ici pour plus de concision, les phrases de contexte sont omises.
8. Plus précisément, nous utilisons le modèle `bert-base-uncased` de HuggingFace (Wolf *et al.*, 2020). Il est composé de 12 couches, avec une taille de représentation cachée de 768 et 12 têtes d'attention, totalisant environ 110 millions de paramètres.

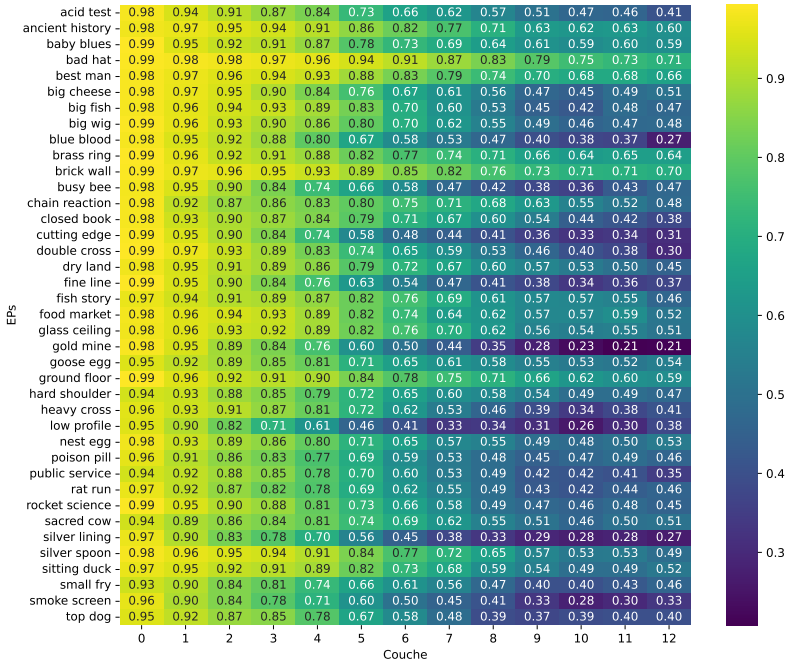

FIGURE 1 – [BERT] Similarités cosinus entre les représentations C et NC des EPs à chaque couche.

Méthode d'évaluation Bien que notre corpus permette une comparaison directe entre les représentations d'expressions polylexicales dans leurs lectures compositionnelles et idiomatiques, sa taille relativement restreinte — tant en nombre d'exemples qu'en diversité d'EPs — ne permet pas l'application de techniques d'analyse standard basées sur des probes supervisées. En effet, les méthodes de *probing* consistent généralement à entraîner un classifieur simple (souvent linéaire) sur les représentations internes d'un modèle afin de prédire une propriété linguistique donnée. Pour être robustes, ces approches requièrent des ensembles de données importants et bien équilibrés, ce qui n'est pas le cas ici : le nombre limité d'expressions, la variabilité sémantique entre les items, et la présence de seulement quelques occurrences par sens ne permettent pas de garantir une généralisation fiable des classifieurs appris.

Plutôt que de chercher à prédire explicitement la lecture d'une expression à partir de ses représentations, nous optons pour une approche exploratoire centrée sur l'analyse structurelle des encodages. Plus précisément, nous examinons comment les représentations des EPs évoluent au fil des couches de BERT, en nous appuyant sur des mesures de distance et de regroupement (*clustering*). Cette stratégie nous permet d'évaluer dans quelle mesure le modèle encode spontanément la distinction sémantique entre lectures idiomatiques et compositionnelles, sans recourir à une supervision externe.

Similarité entre les représentations C et NC Notre première expérience compare les représentations des EPs selon leurs significations. Nous construisons, pour chaque couche, une représentations de chaque usage (C ou NC) de chaque EP en calculant la moyenne des représentations de cet EP pour cet usage. Nous mesurons la similarité cosinus entre les deux centroïdes obtenus pour chaque EP.

Les résultats (Figure 1) révèlent que la similarité cosinus entre les représentations C et NC dans la dernière couche est faible (min = 0,2 ; max = 0,7 ; médiane = 0,47), ce qui démontre la capacité de BERT à différencier ces significations. Cependant, la variation des scores de similarité indique un encodage non homogène, probablement influencé par le degré de compositionnalité propre à chaque

EP	Couche												
	0	**1**	**2**	**3**	4	5	6	7	8	9	10	11	12
acid test	68	**100**	**100**	**100**	100	100	100	100	100	100	100	100	93
ancient history	77	77	81	68	68	68	45	45	77	59	63	90	90
bad hat	70	64	58	52	52	47	52	64	58	64	76	100	100
best man	55	80	95	80	90	75	55	85	95	95	90	90	95
big fish	72	45	72	81	72	72	90	81	81	90	90	90	90
blue blood	47	**100**	88	**100**	100	94	88	88	88	82	88	88	94
brass ring	57	92	84	76	**100**	100	96	92	88	88	88	80	88
brick wall	58	94	82	82	82	76	82	58	76	76	82	76	70
closed book	40	93	93	**100**	93	86	100	93	100	100	100	100	100
cut. edge	55	**100**	**100**	**100**	100	100	100	100	100	100	100	100	100
dry land	76	90	38	95	95	90	85	85	95	100	100	90	95
fine line	76	**100**	92	**100**	100	100	100	100	100	100	100	100	100
grnd. floor	66	**100**	91	79	83	75	91	58	87	100	91	100	100
rock. sci.	60	86	**100**	**100**	100	100	100	100	100	100	100	100	100
silver spoon	46	84	**100**	**100**	92	100	100	100	100	100	100	100	100
% de 100%	0	33	26	47	40	40	40	33	40	53	47	53	47
% de $\geq$90%	0	60	53	53	67	53	60	47	53	67	67	80	87

TABLE 1 – Séparabilité linéaire (%) au travers des couches de BERT.

EP. À l'inverse, dans la couche 0, où les représentations d'entrée sont statiques et sans contexte, les scores de similarité sont élevés (min = 0,9 ; max = 1 ; médiane = 0,98). Mais à mesure que les couches profondes intègrent davantage de contexte, la similarité diminue, ce qui montre la capacité de BERT à raffiner progressivement les distinctions sémantiques entre les significations C et NC.

Une comparaison étendue avec d'autres modèles (BERT-large, Multi-BERT, RoBERTa-base et RoBERTa-large) est présentée en Annexe 5.1. Elle confirme en partie la tendance observée avec BERT-base, tout en révélant des variations intéressantes. Notamment, BERT-large affiche un comportement similaire, avec une distinction maximale entre significations C et NC atteinte dans ses couches intermédiaires, sans gain substantiel par rapport à la version base. Les résultats de Multi-BERT montrent quant à eux une bonne robustesse aux variations d'initialisation, suggérant que les effets observés ne dépendent pas de la graine d'entraînement. Enfin, les deux variantes de RoBERTa produisent des représentations globalement plus dissemblables entre les usages C et NC, avec des similarités parfois négatives dès les premières couches. Ce comportement contraste avec celui de BERT et pourrait refléter une sensibilité différente à la structure contextuelle, probablement liée aux choix de pré-entraînement. Toutefois, il reste à établir si cette séparation marquée correspond réellement à une désambiguïsation sémantique plus fine.

Comparaison directe des représentations Au lieu d'utiliser la distance comme indicateur, notre deuxième expérience examine si les représentations des EC dont le sens est construit de manière compositionnelle et ceux dont le sens est non compositionnel sont linéairement séparables. Pour 15 EPs ayant au moins trois occurrences par signification, nous extrayons leurs représentations, les annotons comme C ou NC, puis entraînons un classifieur linéaire [9] afin d'évaluer s'ils sont linéairement séparables [10]. Une classification parfaite suggère que BERT « organise » distinctement les représentations C et NC.

9. Nous avons choisi un perceptron car il n'est pas régularisé. L'implémentation utilisée provient de la bibliothèque sklearn (Pedregosa *et al.*, 2011).

10. L'objectif est de déterminer si les représentations des EPs sont distinctes et si un classifieur linéaire peut les différencier efficacement, sans chercher à tester sa capacité de généralisation (faute de données suffisantes). Nous avons donc utilisé l'ensemble des données pour l'apprentissage, sans considérer d'ensemble de test.

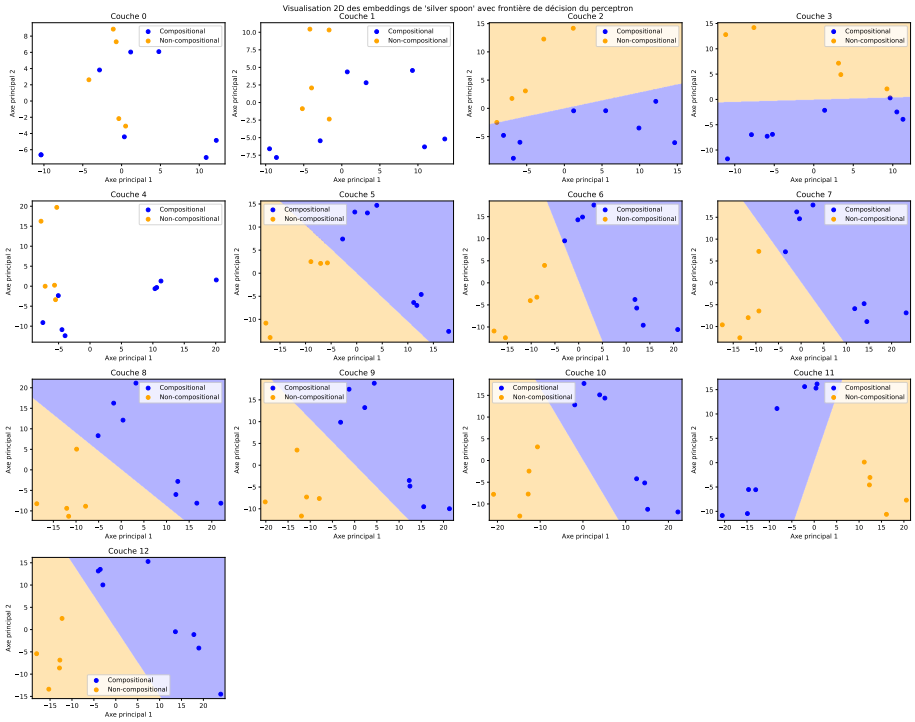

FIGURE 2 – Visualisation de la séparation linéaire entre les occurrences C et NC de *silver spoon*.

Les résultats, résumés à le Tableau 1 indiquent qu'à la dernière couche, 87% des EPs sont presque parfaitement séparables, ce qui témoigne de la capacité de `BERT` à distinguer les significations C et NC. Par ailleurs, parmi les EPs atteignant la séparabilité linéaire ou presque (non grisés dans le tableau), 73% y parviennent dès les premières couches (1-3) (en gras). La Figure 2 illustre ce phénomène avec une projection PCA des représentations de *silver spoon* et la frontière de décision correspondante : dès la couche 2, le modèle parvient à séparer les significations C et NC, bien qu'une légère dégradation apparaisse à la couche 4.

Ces résultats confirment que `BERT` différencie les usages C et NC des EPs, la séparabilité linéaire émergeant fréquemment dès les premières couches. Une analyse comparative avec d'autres modèles (`BERT-large`, `Multi-BERT`, `RoBERTa-base` et `RoBERTa-large`), présentée en Annexe 5.2, révèle que `BERT-large` montre des performances très proches de `BERT-base`, confirmant que l'augmentation de profondeur ne renforce pas nécessairement la séparation des usages. `Multi-BERT` affiche une bonne robustesse aux conditions d'entraînement : plusieurs EPs deviennent linéairement séparables et le restent jusqu'à la dernière couche, tandis que d'autres ne le deviennent jamais, ou seulement de façon transitoire. En revanche, les résultats pour `RoBERTa` sont plus ambigus : bien que la majorité des couches permettent une séparation linéaire, peu d'expressions apparaissent clairement séparées en sortie de modèle. Cela suggère une organisation différente de l'espace représentationnel, dont les implications en termes de désambiguïsation restent à explorer.

Cohésion entre les tokens des groupes idiomatiques Les premières expériences montrent que `BERT` différencie bien les significations C et NC grâce à ses représentations contextualisées. Mais traite-t-il les séquences NC comme des unités lexicales plus cohésives, à l'instar des humains (Bhattasali *et al.*, 2019) ? Pour tester cette hypothèse, nous nous appuyons sur l'intuition suivante : si une EP NC est perçue comme une entité indivisible, on s'attendrait à ce que ses composants aient des

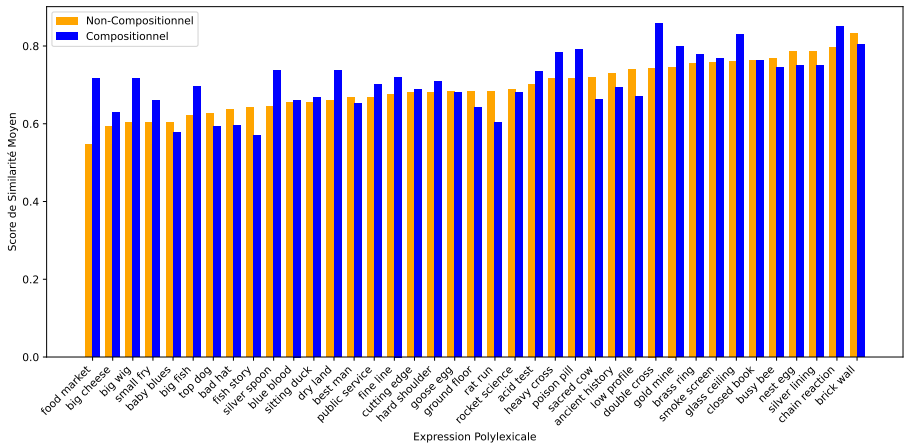

FIGURE 3 – Scores de similarité moyennés entre les composants des EPs en contextes C et NC

représentations plus proches que dans un contexte C.

Pour vérifier cette hypothèse, nous avons mesuré la similarité cosinus entre les représentations des deux mots formant chaque EP, en contextes C et NC, en nous basant sur la couche 12 du modèle, qui est la représentation typiquement utilisée pour les tâches en aval. Nous avons ensuite calculé la similarité moyenne de chaque EP, pondérée par le nombre d'occurrences dans chaque groupe. Cette méthodologie suit celle de Soler *et al.* (2024), qui ont analysé la similarité inter-mots pour évaluer la qualité des représentations idiomatiques.

Les résultats ne révèlent pas d'augmentation notable de la cohésion lexicale en contexte NC (similarité cosinus moyenne : 0,71 en C, 0,7 en NC ; Figure 3). Cela suggère que BERT ne rapproche pas davantage les composants d'une expression idiomatique qu'il ne le fait en contexte compositionnel. De plus, son mécanisme de contextualisation entraîne une similarité élevée entre tous les mots d'une phrase (similarité moyenne de 0,53 pour tous les bigrammes en couche finale), indiquant que la forte proximité observée entre les composants des EPs ne relève pas d'un traitement spécifique des expressions idiomatiques, mais plutôt d'un effet général de l'auto-attention : l'anisotropie des représentations (Ethayarajh, 2019).

En résumé, BERT ne semble pas capturer la cohésion interne des expressions idiomatiques via la proximité de leurs composants dans l'espace des plongements lexicaux. Sa distinction entre C et NC repose davantage sur le contexte global que sur la structure interne des EPs.

4 Discussion et conclusions

Cet article examine comment les transformeurs, en particulier BERT, encodent les syntagmes nominaux non-compositionnels (NC) et leurs équivalents compositionnels (C), en se concentrant sur les expressions idiomatiques qui posent un défi aux modèles de langue. En effet, ces derniers reposent sur des mécanismes de construction du sens qui privilégient l'interprétation des expressions à partir de la combinaison de leurs constituants, ce qui complique la modélisation des usages idiomatiques, dont le sens ne peut être déduit compositionnellement. Nous avons mené trois expériences visant à déterminer si (1) BERT encode différemment les significations C et NC, (2) ces encodages sont linéairement séparables et (3) les expressions NC forment des unités lexicales plus cohésives.

Nos résultats montrent que BERT différencie systématiquement les significations C et NC. Dans la première expérience, les représentations NC affichent une similarité systématiquement plus faible avec leurs équivalents C, en particulier dans les couches profondes, ce qui souligne la capacité de BERT à distinguer les significations idiomatiques des significations littérales. Certaines expressions (p.ex. *best man* [11]) conservent une forte similarité entre C et NC à travers les couches, tandis que d'autres (p.ex. *gold mine* [12]) montrent une divergence marquée, illustrant la variabilité du modèle dans son traitement de la compositionnalité. Ces résultats suggèrent que, bien que BERT soit intrinsèquement orienté vers la compositionnalité, il capture efficacement des significations non compositionnelles sans supervision explicite.

La deuxième expérience renforce cette distinction en montrant que les représentations C et NC sont souvent linéairement séparables, avec une séparation détectable dès la couche 1 pour un tiers des expressions et qui se maintient dans les couches plus profondes. Toutefois, cette séparabilité varie selon les expressions : par exemple, *big fish* [13] présente une séparation plus instable, alors que *cutting edge* [14] est parfaitement séparé dès la première couche. Ces résultats confirment que la compositionalité n'est pas une opposition binaire, mais un continuum, et que le traitement des expressions idiomatiques par BERT est à la fois flexible et contextuel.

En revanche, la troisième expérience ne met en évidence aucune augmentation significative de la cohésion lexicale des syntagmes NC par rapport aux syntagmes C. Cela suggère que BERT ne « compresse » pas les significations idiomatiques en une représentation unique et indivisible, mais qu'il distingue les usages idiomatiques des usages littéraux en s'appuyant sur des représentations contextuelles plus larges.

Dans l'ensemble, BERT démontre sa capacité à différencier les significations idiomatiques et littérales en traitant les expressions idiomatiques comme des séquences polysémiques. Son approche, à la fois contextuelle et dynamique, oscille entre un traitement compositionnel et un traitement plus global.

La comparaison avec d'autres architectures transformeur — en particulier BERT-large, MultiBERT, RoBERTa-base et RoBERTa-large — permet d'évaluer dans quelle mesure les observations faites avec BERT-base sont spécifiques à ce modèle. Alors que BERT-large et MultiBERT montrent des tendances proches de BERT-base, les deux variantes de RoBERTa présentent des profils sensiblement différents, en particulier dans les premières couches. Ces résultats montrent que, malgré certaines variations entre modèles, plusieurs tendances observées avec BERT-base — en particulier la capacité à distinguer les usages idiomatiques et compositionnels dans les couches intermédiaires à profondes — se retrouvent également dans BERT-large et MultiBERT, et dans une certaine mesure dans RoBERTa. Cela suggère l'existence de dynamiques de représentation relativement robustes à travers différentes architectures et conditions d'entraînement, ouvrant la voie à des généralisations prudentes mais fondées.

Ces résultats ouvrent des perspectives pour les recherches futures. Celles-ci pourraient explorer les mécanismes d'attention qui sous-tendent les différences d'encodage entre expressions compositionnelles et non compositionnelles, afin de mieux comprendre comment les modèles capturent les variations de sens induites par le contexte. Étendre l'analyse à d'autres langues et modèles permettrait également d'évaluer la robustesse et la généralisation des effets observés. Enfin, la conception ou

11. Littéralement « meilleur homme », pouvant aussi signifier « le témoin du marié ».
12. Littéralement « mine d'or », pouvant aussi signifier « source extrêmement riche en ressources, en opportunités ou en informations précieuses ».
13. Littéralement « gros poisson », pouvant aussi signifier « personne influente ou importante ».
14. Littéralement « bord tranchant », pouvant aussi signifier « à la pointe de l'innovation ».

l'utilisation de jeux de données intégrant des gradients plus fins de compositionnalité offrirait un levier précieux pour mieux comprendre le continuum entre langage littéral et langage idiomatique.

Remerciements

Cette recherche a été soutenue financièrement par l'*Agence Nationale de la Recherch*e (projet COMPO, ANR-23-CE23-0031-01).

Références

BALDWIN T. & KIM S. N. (2010). Multiword expressions. In *Handbook of Natural Language Processing*.

BHATTASALI S., FABRE M., LUH W.-M., AL SAIED H., CONSTANT M., PALLIER C., BRENNAN J. R., SPRENG R. N. & HALE J. (2019). Localising memory retrieval and syntactic composition : an fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, **34**(4), 491–510. DOI : 10.1080/23273798.2018.1518533.

CONSTANT M., ERYIĞIT G., MONTI J., VAN DER PLAS L., RAMISCH C., ROSNER M. & TODIRASCU A. (2017). Multiword Expression Processing : A Survey. *Computational Linguistics*, **43**(4), 837–892. DOI : 10.1162/COLI_a_00302.

DANKERS V., LUCAS C. & TITOV I. (2022). Can Transformer be Too Compositional ? Analysing Idiom Processing in Neural Machine Translation. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3608–3626, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.252.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

ETHAYARAJH K. (2019). How contextual are contextualized word representations ? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 55–65, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1006.

HASHEMPOUR R. & VILLAVICENCIO A. (2020). Leveraging Contextual Embeddings and Idiom Principle for Detecting Idiomaticity in Potentially Idiomatic Expressions. In M. ZOCK, E. CHERSONI, A. LENCI & E. SANTUS, Éds., *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, p. 72–80, Online : Association for Computational Linguistics.

HUPKES D. (2020). *Hierarchy and interpretability in neural models of language processing*. doctoral, University of Amsterdam.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2019). Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

LIU E. & NEUBIG G. (2022). Are representations built from the ground up ? An empirical examination of local composition in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 9053–9073, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.617.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *CoRR*, **abs/1907.11692**.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality.

MITCHELL J. & LAPATA M. (2010). Composition in distributional models of semantics. *Cognitive science*, **34 8**, 1388–429.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

SAVARY A., CORDEIRO S., LICHTE T., RAMISCH C., IÑURRIETA U. & GIOULI V. (2019). Literal Occurrences of Multiword Expressions : Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical Linguistics*. DOI : 10.2478/pralin-2019-0001, HAL : hal-02106263.

SELLAM T., YADLOWSKY S., WEI J., SAPHRA N., D'AMOUR A., LINZEN T., BASTINGS J., TURC I., EISENSTEIN J., DAS D., TENNEY I. & PAVLICK E. (2021). The multiberts : BERT reproductions for robustness analysis. *CoRR*, **abs/2106.16163**.

SOLER A. G., LABEAU M. & CLAVEL C. (2024). The Impact of Word Splitting on the Semantic Content of Contextualized Word Representations. arXiv :2402.14616 [cs], DOI : 10.48550/arXiv.2402.14616.

SZABÓ Z. G. (2020). Compositionality. In E. N. ZALTA & U. NODELMAN, Éds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 édition.

TAYYAR MADABUSHI H., GOW-SMITH E., SCARTON C. & VILLAVICENCIO A. (2021). ASTitchInLanguageModels : Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 3464–3477, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-emnlp.294.

TIAN Y., JAMES I. & SON H. (2023). How are idioms processed inside transformer language models ? In A. PALMER & J. CAMACHO-COLLADOS, Éds., *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, p. 174–179, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.starsem-1.16.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention Is All You Need. arXiv :1706.03762 [cs], DOI : 10.48550/arXiv.1706.03762.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers : State-of-the-art natural language processing. In Q. LIU & D. SCHLANGEN, Éds., *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

ZENG Z. & BHAT S. (2022). Getting BART to ride the idiomatic train : Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, **10**, 1120–1137. DOI : 10.1162/tacl_a_00510.

5 Annexes

5.1 Comparaison avec d'autres modèles transformeurs

Afin de compléter notre analyse centrée sur `BERT`, nous avons étendu l'étude à la version *large* de `BERT`, à deux variantes du modèle `RoBERTa` (`base` et `base`) (Liu *et al.*, 2019), ainsi qu'à `multiBERT` (Sellam *et al.*, 2021), un ensemble de modèles `BERT` entraînés à partir de graines différentes. Ces modèles partagent la même architecture de base que `BERT`, mais diffèrent notamment par la taille et la nature des données d'entraînement (`BERT-large`, `RoBERTa`), ou les conditions d'initialisation (`multiBERT`). Ces différences permettent d'observer dans quelle mesure les représentations des expressions polylexicales sont sensibles aux choix de conception du modèle. Pour rester cohérents avec l'expérience principale, nous limitons l'analyse aux EPs tokenisées en exactement deux sous-unités.

La Figure 4 montre que les performances de `BERT-large` sont globalement similaires à celles de `BERT-base`. Dans les deux cas, la similarité entre les représentations compositionnelles et idiomatiques des EPs est relativement élevée dans les couches inférieures, ce qui suggère une représentation encore peu influencée par le contexte. La Figure 4 révèle que les couches les plus profondes de `BERT-large` — en particulier autour des couches 13 à 20 — sont celles où la similarité entre les représentations compositionnelles et idiomatiques est la plus faible, ce qui indique une distinction sémantique plus importante du modèle en fonction du contexte. Toutefois, ces scores restent très proches de ceux observés dans les couches profondes (8 à 12) de `BERT-base`. Cela suggère que l'augmentation de la profondeur n'apporte pas de gain substantiel en termes de désambiguïsation sémantique dans ce cadre précis.

La Figure 5 montre les résultats obtenus avec `multiBERT` : on observe une certaine stabilité des scores de similarité à travers les différentes initialisations, et des tendances similaires à celles de `BERT`, bien que la désambiguïsation sémantique ne soit pas aussi marquée (min = 0.47). Cette robustesse suggère que l'encodage des EPs par `BERT` n'est pas fortement dépendant de la graine d'entraînement, ce qui conforte la fiabilité des résultats obtenus dans notre expérience principale.

La Figure 6 présente les résultats pour `RoBERTa-base`. Contrairement à `BERT`, les scores de similarité sont majoritairement faibles, parfois même négatifs, et ce dès les premières couches. Ce contraste marqué pourrait s'expliquer par l'absence de la tâche de *Next Sentence Prediction* dans l'entraînement de `RoBERTa`, un facteur potentiellement important pour les expressions sensibles au contexte. La version `RoBERTa-large` suit des tendances similaires, malgré une légère amélioration des scores dans les couches profondes.

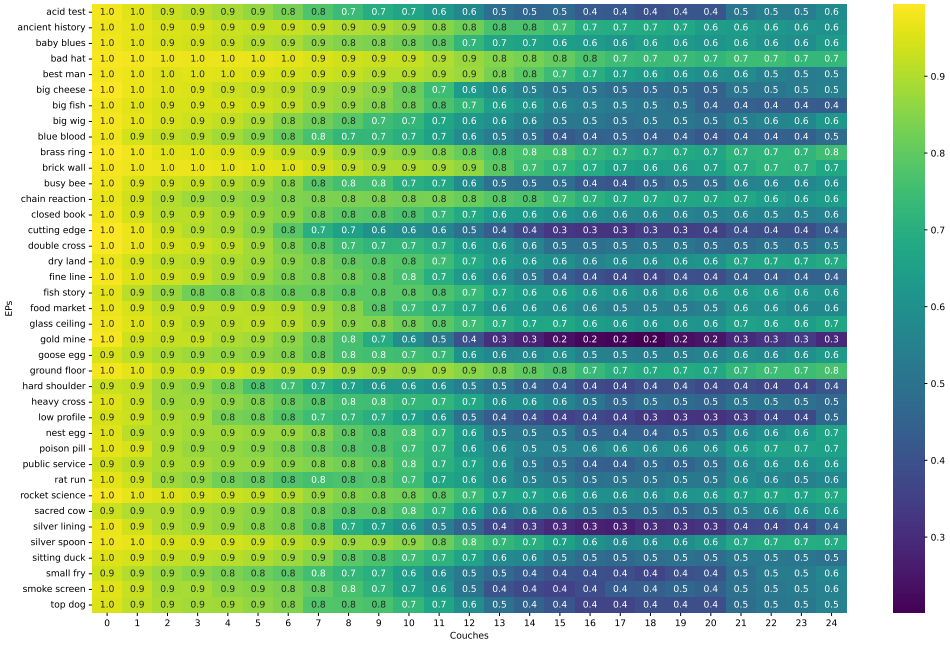

FIGURE 4 – [BERT-large] Similarités cosinus entre les représentations C et NC des EPs à chaque couche.

.

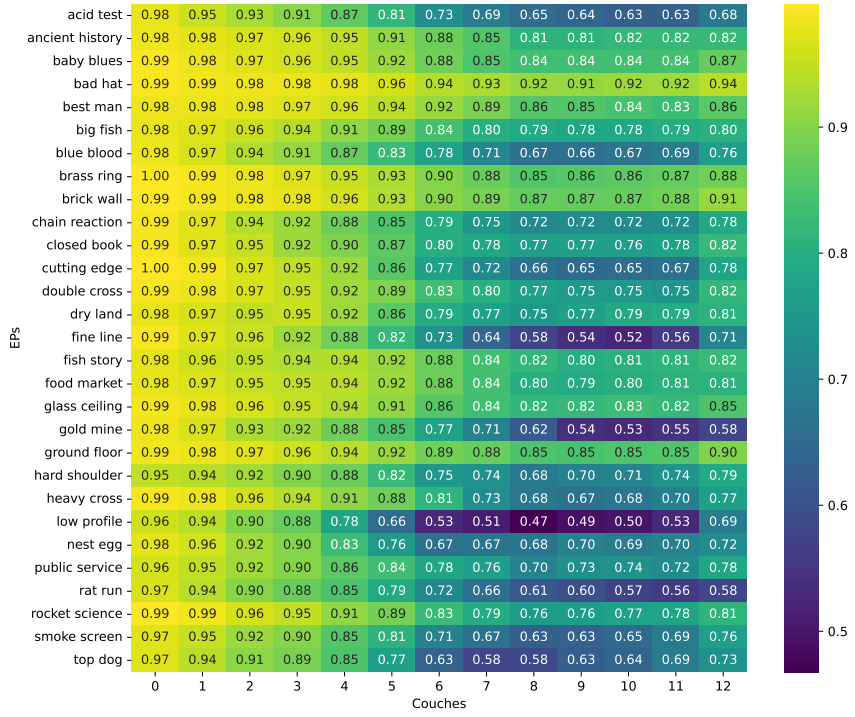

FIGURE 5 – [MultiBERT] Similarités cosinus entre les représentations C et NC des EPs à chaque couche.

.

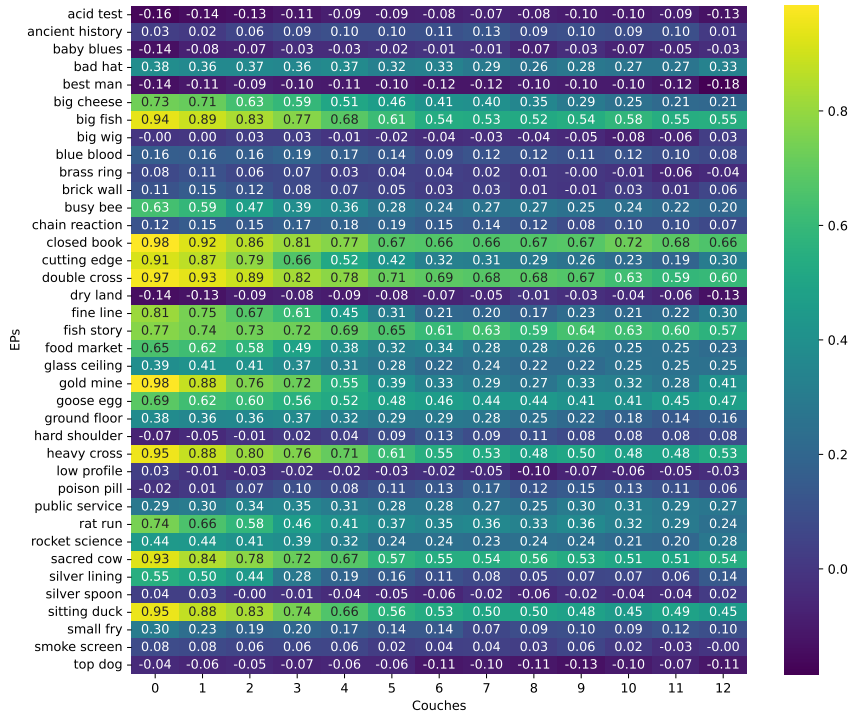

FIGURE 6 – [RoBERTa-base] Similarités cosinus entre les représentations C et NC des EPs à chaque couche.

5.2 Séparabilité linéaire des représentations — comparaison inter-modèles

En complément des résultats obtenus avec BERT-base, nous avons évalué la séparabilité linéaire des représentations C et NC dans plusieurs autres modèles transformeurs : BERT-large, Multi-BERT, RoBERTa-base et RoBERTa-large. L'objectif est d'examiner dans quelle mesure les différences architecturales ou de pré-entraînement influencent la capacité à distinguer les usages compositionnels et non compositionnels des EPs. Comme pour l'étude principale, nous considérons ici les EPs pour lesquelles il y a au moins trois occurrences par sens.

BERT-large. Le modèle BERT-large montre une dynamique très similaire à celle de BERT-base. Le Tableau 2 montre que dans la majorité des cas (10 expressions sur 17), la séparabilité linéaire entre les représentations compositionnelles et non compositionnelles est atteinte et maintenue jusqu'à la dernière couche du modèle. Pour 4 expressions, cette séparabilité est observée à un moment donné mais n'est pas conservée en sortie, tandis que 3 expressions ne deviennent jamais séparables.

Multi-BERT. Les résultats agrégés sur les variantes de Multi-BERT indiquent une bonne robustesse aux variations d'initialisation : le tableau 3 montre que dans 50% des cas, les EPs deviennent linéairement séparables à un certain point du modèle et conservent cette séparabilité jusqu'à la dernière couche. Pour une EP, la séparabilité est atteinte de façon transitoire mais perdue à la sortie. Dans 37,7% des cas, les représentations ne sont jamais devenues linéairement séparables.

RoBERTa. Les deux variantes de RoBERTa atteignent des taux élevés de séparabilité linéaire dans les couches intermédiaires, mais les résultats en sortie de modèle sont moins convaincants. Comme le montre le tableau 4, à l'exception de l'expression *cutting edge*, aucune EP ne montre une séparation en sortie du modèle. Cela suggère que, bien que RoBERTa encode une distinction entre les usages

	Couche																							
EP	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
acid test	88	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	94	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
ancient history	77	82	73	82	82	77	77	77	82	82	82	68	68	73	68	41	68	50	86	91	91	91	91	77
baby blues	75	70	80	**100**	**100**	90	**100**	90	**100**	**100**	65	**100**	**100**	65	**100**	85	85	90	75	85	85	70	80	80
bad hat	59	53	71	65	59	71	88	**100**	**100**	**100**	**100**	**100**	76	82	53	88	76	82	76	76	76	65	76	59
best man	75	85	80	95	95	95	85	90	90	95	95	70	85	80	90	90	90	90	95	95	**100**	**100**	**100**	**100**
big fish	73	64	82	55	82	**100**	**100**	**100**	**100**	91	91	91	82	73	82	82	73	91	**100**	**100**	**100**	**100**	**100**	**100**
blue blood	82	**100**	**100**	94	**100**	76	94	76	94	82	94	94	94	94	94	94	88	88	88	94	94	88	**100**	88
brass ring	81	96	85	92	85	77	96	92	88	88	92	85	88	88	88	88	92	81	85	85	81	88	81	69
brick wall	76	88	82	88	94	82	76	71	47	88	82	82	88	88	82	88	76	88	88	82	82	94	71	71
closed book	80	93	87	93	93	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	93	93	93	**100**	**100**	87	73	80	73
cutting edge	89	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
dry land	76	90	95	90	90	95	95	95	95	95	95	90	95	95	95	90	95	90	95	95	81	95	95	**100**
fine line	85	69	85	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
gold mine	88	81	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
ground floor	83	92	96	83	88	75	83	92	83	**100**	83	75	75	83	79	79	75	50	79	79	75	79	38	58
rocket science	80	87	93	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	93	**100**	93	93	**100**	**100**	**100**	**100**
silver spoon	85	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	92	**100**	92	92	**100**
% de 100%	0	24	29	41	47	47	53	53	59	59	47	47	47	41	47	35	29	35	41	35	47	41	47	53
% de $\geq$90%	0	47	47	71	71	65	71	82	76	76	76	71	59	53	65	59	59	65	59	71	59	65	65	53

TABLE 2 – [BERT-large] Séparabilité linéaire (%) au travers des couches du modèle.

	Couche											
EP	0	1	2	3	4	5	6	7	8	9	10	11
acid test	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
ancient history	82	73	77	82	77	73	82	45	77	55	64	82
baby blues	65	82	65	59	71	71	76	71	82	65	82	71
bad hat	44	50	44	56	50	50	62	44	31	31	31	56
best man	68	68	68	63	89	74	89	84	95	**100**	**100**	**100**
big fish	73	82	82	82	82	82	91	73	91	91	91	**100**
blue blood	76	**100**	**100**	**100**	76	94	94	94	88	94	94	82
brass ring	58	46	92	92	96	85	92	92	88	88	88	92
brick wall	65	47	71	88	82	82	82	82	82	76	76	76
closed book	80	87	93	93	93	87	93	80	87	60	93	80
cutting edge	56	67	89	89	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
dry land	90	71	67	90	95	**100**	**100**	**100**	**100**	95	33	**100**
fine line	85	85	92	92	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
gold mine	94	94	**100**	94	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
ground floor	78	52	65	91	87	78	87	74	91	87	83	83
rocket science	67	87	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
% de 100%	6	12	25	19	31	38	38	38	38	38	38	50
% de $\geq$90%	19	19	44	56	50	44	62	50	56	56	56	56

TABLE 3 – [MultiBERT] Séparabilité linéaire (%) au travers des couches du modèle.

	Couche											
EP	0	1	2	3	4	5	6	7	8	9	10	11
acid test	88	75	**100**	62	44	44	75	75	81	**100**	**100**	75
ancient history	77	68	82	73	77	86	82	86	64	73	77	73
bad hat	53	67	73	73	27	80	73	73	73	87	87	53
best man	84	63	95	95	84	95	89	84	84	95	79	84
big fish	45	45	64	64	64	82	82	91	64	91	64	91
blue blood	94	94	94	94	**100**	94	94	88	94	94	76	76
brass ring	72	72	80	80	88	80	92	88	84	76	84	56
brick wall	93	93	93	93	93	93	80	93	93	93	93	93
closed book	80	87	87	93	**100**	**100**	**100**	**100**	**100**	93	80	93
cutting edge	44	56	78	94	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
dry land	81	81	86	48	95	43	81	95	95	81	86	95
fine line	92	85	77	69	85	62	69	85	85	85	**100**	85
ground floor	91	91	91	91	91	91	91	70	83	**100**	91	91
rocket science	80	87	87	93	87	80	80	87	93	80	67	60
silver spoon	85	92	69	77	62	85	85	69	62	54	31	62
% de 100%	0	0	7	0	20	13	13	13	13	20	20	7
% de $\geq$90%	27	27	33	47	40	40	33	33	40	53	33	40

TABLE 4 – [RoBERTa-base] Séparabilité linéaire (%) au travers des couches du modèle.

C et NC, cette distinction n'est pas exploitée de façon aussi stable ou lisible en sortie. Des analyses supplémentaires seront nécessaires pour comprendre cette différence structurelle par rapport à BERT.