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

Exploration de la modalité en français parlé et écrit

Anna Colli[1] Delphine Battistelli[1]
(1) Laboratoire MoDyCo (UMR 7114 Université Paris Nanterre-CNRS)
acolli@parisnanterre.fr, dbattist@parisnanterre.fr

RÉSUMÉ

Dans cet article, nous présentons une méthodologie pour comparer entre eux les profils modaux de corpus en français. Nous montrons quelles différences émergent ou non entre l'écrit et l'oral et pointons l'importance et la place des marqueurs polysémiques dans les deux cas. L'analyse de la polysémie du verbe *pouvoir* retient notre attention dans la mesure où ce verbe s'avère être un marqueur très présent dans l'ensemble des corpus.

ABSTRACT

Exploring modality in spoken and written French

In this article, we present a methodology for comparing modal profiles in French corpora. We show differences between written and spoken French and we point out the importance and place of polysemous markers in both cases. The analysis of the polysemy of the verb *pouvoir* attracts our attention as this verb turns out to be a very present marker in all the corpora.

MOTS-CLÉS : modalité, polysémie, profil modal, oral, écrit.

KEYWORDS: modality, polysemy, modal profile, spoken data, written data.

1 Introduction

La catégorie sémantique de la modalité désigne - du moins dans une conception large principalement initiée dans les travaux de (Bally, 1932) et poursuivie notamment dans ceux de (Gosselin, 2010) - l'ensemble des procédés linguistiques qui renvoient à l'expression de l'attitude de l'énonciateur vis-à-vis d'un contenu propositionnel. Dans le travail que nous présentons ici, nous nous focalisons sur les marqueurs modaux uniquement lexicaux. Notre approche vise à combiner deux perspectives pour l'analyse de la modalité dans les textes : une perspective de linguistique de corpus, axée sur l'exploration de corpus de nature différente ; et une perspective de TAL, en développant certains outils utiles à cette exploration.
La finalité est alors d'analyser quels marqueurs sont les plus fréquents dans ces différents corpus et de mesurer à quel point ce sont les mêmes. Nous nous interrogeons également sur le fait que, dans le cas de marqueurs polysémiques, ce soient ou non les mêmes valeurs modales qui seront privilégiées quel que soit le corpus, notamment et surtout dans l'opposition oral/écrit.

Après un bref état de l'art des travaux en TAL (Section 2), nous décrivons les principaux éléments du cadre théorique dans lequel nous nous situons ainsi que la ressource lexicale que nous avons employée, elle-même adossée sur ce cadre théorique (Section 3). Nous montrons ensuite les résultats obtenus suite à l'utilisation de cette ressource dans l'analyse de ce que nous nommons le "profil modal" d'un corpus (Section 4). Nous détaillons dans la Section 5 le cas de l'analyse du verbe polysémique

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 376–389.

pouvoir et montrons comment sa désambiguisation automatique accomplie par un classifieur que nous avons développé permet d'ajuster si nécessaire les calculs du profil modal d'un corpus.

2 Repérer et catégoriser les marqueurs modaux

Prenons les exemples 1., 2. et 3. (issus de trois corpus distincts sur lesquels nous reviendrons) dans lesquels les marqueurs modaux lexicaux ont été soulignés.

1. *C'est une meuf, qui déteste les meufs.* (extrait de TWEETS) : évaluation négative
2. *Je me suis précipitée sur sur Google sur l'appli Le Monde et tout pour essayer de capter quoi* (extrait de 13_11) : désir *et* incertitude
3. *Si vraiment je peux pas le faire je le ferai pas* (extrait de ESLO_ENT) : capacité *ou* permission

Dans (1), le locuteur rapporte une évaluation négative portée par l'emploi du lexème modal *détester* ; ceci est réalisé de manière univoque, c'est-à-dire qu'il n'y a qu'une valeur modale possible associée au marqueur. Dans (2) et (3), la situation est différente : il y a deux valeurs modales associées à un même marqueur modal (resp. *essayer* et *pouvoir*) : dans (2), les valeurs modales désir *et* incertitude sont associés à *essayer* ; dans (3), seul l'ajout d'un contexte comme « *parce que je manque d'expertise* » ou « *parce que mon chef l'a interdit* » permettrait d'orienter le choix vers une valeur de capacité ou de permission, respectivement. Dans ce dernier cas, on parle de polysémie d'un marqueur et d'ambiguïté entre valeurs modales.
En TAL, comme du reste en linguistique, la diversité des conceptions théoriques de la modalité empêche l'émergence d'un standard commun comme en témoigne les nombreux schémas d'annotation proposés e.g. (Ruppenhofer & Rehbein, 2012; Rubinstein *et al.*, 2013; Nissim *et al.*, 2013; Pyatkin *et al.*, 2021). En revanche, tous ces travaux s'accordent sur le fait que la polysémie des marqueurs modaux en représente un des principaux défis. En effet, il faut acquérir des informations supplémentaires sur leur contexte linguistique (voire extra-linguistique) afin d'en déterminer la valeur modale (on parle de désambiguïsation). On distinguera alors deux types de travaux selon leur façon d'aborder la polysémie. Il y a ceux qui se focalisent sur la désambiguisation automatique, mais en général ce sont seulement quelques marqueurs en particulier qui sont visés, e.g. les auxiliaires *can*, *could*, *may*, *must* dans un corpus de dépêches en anglais (Ruppenhofer & Rehbein, 2012). Il y a d'autres travaux comme (Baker *et al.*, 2010) et (Battistelli & Étienne, 2025), à l'origine de deux ressources lexicales respectivement pour l'anglais et pour le français, qui font référence à ce phénomène en listant les différentes valeurs possibles mais ne proposent pas de procédures de désambiguisation. La plupart des travaux en TAL se concentre sur l'anglais. Parmi les rares travaux sur le français (Nissim *et al.*, 2013; Battistelli & Étienne, 2025), on notera que, même s'il s'agit de cadres théoriques différents, ils partagent les mêmes catégories modales principales (i.e. épistémique, déontique, axiologique et appréciative). Du côté de la linguistique de corpus, mais uniquement pour l'anglais (e.g.(Rizvic-Eminovic & Šukalić, 2019)), on note par ailleurs que des travaux se sont intéressés à l'analyse de la distribution de certains marqueurs modaux (en réalité, les verbes modaux seuls) dans des textes de nature différente (oral vs écrit académique). Dans l'ensemble des travaux menés en TAL, il apparaît ainsi qu'il n'y a pas eu jusqu'ici de travaux menés en vue d'observer la manière dont la modalité s'exprime différemment selon que l'on se situe à l'oral et à l'écrit, en français, à la fois en termes de marqueurs lexicaux privilégiés et en termes de catégories modales privilégiées. C'est la voie que nous avons commencée à explorer et que nous décrivons ici.

3 Cadre théorique adopté

La catégorie sémantique de la modalité est caractérisée par une grande diversité de définitions et d'organisation interne du domaine. Nous nous appuyons pour notre part sur le cadre théorique proposé par (Gosselin, 2010) qui organise la modalité selon 6 catégories : aléthique, épistémique, appréciative, axiologique, boulique et déontique. La modalité aléthique concerne les propositions présentées comme des vérités objectives alors que l'épistémique, l'appréciative et l'axiologique regroupent les marqueurs qui indiquent une évaluation subjective du locuteur basée, respectivement, sur ses croyances, sa désidérabilité/indésidérabilité ou de lois morales ou éthiques. La modalité boulique concerne les désirs, les volontés ou les souhaits et la modalité déontique l'obligatoire, l'interdit, le permis ou le facultatif. Gosselin adopte une conception large de la modalité qui se démarque des conceptions dites "étroites" (lesquelles sont bien souvent retenues en TAL) qui réduisent l'ensemble des marqueurs modaux aux verbes modaux (*devoir*, *pouvoir*, *vouloir*) - voir notamment (Barbet & de Saussure, 2012). Dans l'approche de Gosselin, il n'existe pas de classe homogène de marqueurs modaux partageant un même statut morpho-syntaxique ; plus encore, tout lexème est porteur d'au moins une modalité qui est, par défaut, aléthique.

Pour notre travail d'analyse du profil modal d'un corpus, nous avons employé le processeur ModalE (Battistelli & Étienne, 2025) [1], une ressource lexicale de 631 marqueurs construite à partir des exemples et des définitions issus de l'ouvrage (Gosselin, 2010) et enrichie selon des critères sémantiques comme, par exemple, ceux de synonymie ou d'appartenance à un même champ lexical. ModalE repère et classe les marqueurs modaux d'un texte donné en entrée sur la base des 6 catégories modales vues ci-dessus. Il s'agit d'un outil modulable à l'aide de filtres qui permettent de sélectionner la ou les catégories modales à analyser. En outre, il est possible de filtrer aussi d'autres paramètres comme, par exemple, le caractère dénoté ou non des marqueurs modaux [2]. Le filtrage permet, par exemple, d'exclure les marqueurs qui expriment exclusivement des modalités aléthiques non-dénotées. Il est à noter que, classiquement, en TAL (Ruppenhofer & Rehbein, 2012; Rubinstein *et al.*, 2013; Pyatkin *et al.*, 2021) et en lien avec la conception étroite de la modalité qui leurs est généralement attachée, ces marqueurs à valeur aléthique ne sont pas retenus dans les tâches d'annotation car considérés comme non-modaux. La ressource ModalE est modulable, c'est à dire que l'on peut choisir d'appliquer des filtres. Dans le cadre d'utilisation de ModalE présentée ici, nous avons ainsi, par exemple, décidé d'exclure de nos analyses les marqueurs qui expriment exclusivement des modalités aléthiques non-dénotées. Le filtrage permettrait aussi de s'intéresser à des marqueurs spécifiques (comme, par exemple, les verbes modaux) ou à des catégories modales spécifiques. [3] Il est à noter que nous avons enrichi la première version de ModalE de manière à ce qu'elle puisse différencier des marqueurs comme dans les exemples 2. (où un marqueur modal dénote plusieurs catégories modales de manière simultanée) et 3. (où un marqueur modal est polysémique, c'est-à-dire qu'il renvoie à une des valeurs en fonction du contexte linguistique) ; ce qui constitue autant de filtres à nouveau possible pour utiliser ModalE. Pour cet enrichissement de ModalE, nous sommes parties plus précisément de la terminologie proposée par (Gosselin, 2010) qui distingue trois types de marqueurs :

— Les marqueurs de type noté **Mono** (dans ModalE, ils sont 426, soit 68% de la ressource) : ce

1. Accessible en ligne ici : https://texttokids.ortolang.fr/semantics/, onglet Modalité

2. Les marqueurs dénotés dénotent les valeurs modales. Par exemple, *possible* et *pouvoir* dénotent la valeur de possibilité de la catégorie aléthique ou *certitude* dénote la valeur de certitude de la catégorie épistémique. Les marqueurs non-dénotés ne dénotent pas une valeur modale. Par exemple, *table*, *frère*, *manger* marquent la catégorie aléthique, sans la dénoter

3. Plus généralement, ce caractère modulable de ModalE permet de comparer nos résultats avec ceux issus d'autres travaux qui reposeraient sur des choix théoriques différents où par exemple certains marqueurs modaux ou certaines valeurs modales ne seraient pas retenus.

sont des marqueurs qui ne relèvent que d'une seule catégorie modale. Par exemple, l'adjectif *injuste* ou le verbe *détester* sont des marqueurs Mono puisqu'ils ne renvoient respectivement qu'à la modalité axiologique et à la modalité appréciative ;

— Les marqueurs de type noté **Mixtes** (dans ModalE, ils sont 175, soit 28% de la ressource) : ce sont des marqueurs qui renvoient, même en contexte, à plusieurs catégories modales. Par exemple, les marqueurs *espérer* et *essayer* marquent dans tout contexte une modalité boulique et une modalité épistémique ;

— Les marqueurs de type noté **Poly** (dans ModalE, ils sont 30, soit 5% de la ressource) : ce sont des marqueurs qui relèvent selon leur contexte, d'au moins deux catégories modales différentes. Par exemple, le verbe *pouvoir* peut marquer la modalité déontique (avec une valeur de permission), la modalité épistémique (avec une valeur d'éventualité) ou aléthique (avec une valeur de capacité) selon le contexte.

Comme nous le verrons plus loin (Section 5), le développement de programmes spécifiques à la désambiguïsation des marqueurs Poly vient compléter ModalE. Cela est indispensable pour appréhender la modalité en corpus car, comme nous allons le voir dans la section suivante, bien que minoritaires dans la ressource ModalE, les marqueurs polysémiques s'avèrent être quantitativement les plus présents dans les corpus.

4 Calculer le profil modal d'un corpus

Notre démarche consiste à calculer ce que nous appelons le "profil modal" d'un corpus. Nous avons commencé par réaliser cette tâche pour deux corpus de données orales et nous l'avons ensuite étendue à deux corpus de données écrites afin d'en dégager les similarités et les différences en termes de marqueurs employés et de catégories exprimées.
Nous définissons le profil modal d'un corpus de textes (ou même d'un texte ou d'une phrase) relativement à trois types d'informations modales : (Info_1) les proportions de catégories modales qui y sont exprimées ; (Info_2) le TOP20 des marqueurs modaux utilisés [4] ; (Info_3) la proportion des marqueurs Poly employés par rapport aux marqueurs des types Mono et Mixte.

4.1 Cas de corpus oraux

Le premier corpus oral, nommé 13_11, est un corpus d'entretiens retranscrits (environ 500.000 tokens), issus de l'Étude 1000 du Programme 13-novembre [5] autour des attentats du 13 novembre 2015 en Ile de France. Le second corpus oral, nommé ici ESLO_ENT, est extrait du corpus ESLO [6]. Il contient 207 entretiens semi-dirigés menés avec les habitants de la ville d'Orléans sur leur quotidien (environ 500.000 tokens).

D'après la ressource ModalE, le corpus 13_11 comporte 414 marqueurs modaux (avec un total de 33.620 occurrences) et le corpus ESLO_ENT comprend 417 marqueurs modaux (avec un total de 32.926 occurrences). Les TOP20 des marqueurs sont disponibles en Annexe A, les TOP5 en Table 1. En premier lieu, nous remarquons que les 20 marqueurs les plus fréquents couvrent 62%

4. Ce choix se base sur la représentativité des TOP20 marqueurs, qui sera mise en lumière en section 4.1
5. https://www.memoire13novembre.fr/
6. http://eslo.huma-num.fr/

	corpus 13_11			corpus ESLO_ENT		
	marqueur	*type*	*catégorie(s)*	*marqueur*	*type*	*catégorie(s)*
1	**dire**	Poly	D ;E	**dire**	Poly	D ;E
2	**savoir**	Poly	AL ;E	**bien**	Poly	AP ;AX
3	**pouvoir**	Poly	AL ;D ;E	**bon**	Poly	AP ;AX
4	**bon**	Poly	AP ;AX	**pouvoir**	Poly	AL ;D ;E
5	**penser**	Mono	E	**savoir**	Poly	AL ;E

TABLE 1 – TOP5 marqueurs modaux dans 13_11 et ESLO_ENT. En gras : les marqueurs en commun entre les deux corpus

des occurrences des marqueurs modaux dans 13_11 et 61% dans ESLO_ENT. Les 394 marqueurs restants pour 13_11 et les 397 restants pour ESLO_ENT couvrent environ 40% des occurrences dans les deux corpus. Cela met en évidence le fait qu'un nombre restreint de marqueurs représente une partie significative des occurrences totales. A noter que pour la suite de la présentation, nous avons choisi de nous appuyer uniquement sur l'analyse des TOP20 dans le calcul des profils modaux des corpus. Détaillons maintenant les profils modaux des deux corpus oraux :

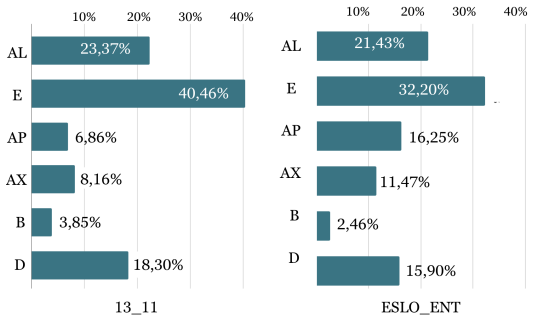

FIGURE 1 – Fréquence normalisée des catégories modales des TOP20 dans les deux corpus oraux

Info_1[catégories modales exprimées] : les catégories les plus fréquentes sont l'épistémique et l'aléthique (Figure 1). Ces catégories sont majoritairement exprimées par des marqueurs Poly et parmi le TOP5 des marqueurs les plus fréquents des deux corpus nous en retrouvons 3 similaires : *dire*, *savoir* et *pouvoir*. Ces résultats montrent déjà que les catégories modales les plus représentées dans les corpus semblent être majoritairement marquées par des marqueurs Poly.
Info_2 [marqueurs modaux employés] : Les deux corpus partagent 75% des marqueurs du TOP20 (Annexe A) et dans le TOP5, il y a même 4 marqueurs identiques (*dire*, *savoir*, *pouvoir*, *bon*), tous de type Poly.
Info_3 [proportion de marqueurs Poly] : alors que les marqueurs Poly représentent seulement 5% de ModalE, ils couvrent 42% (dans 13_11) et 41% (dans ESLO_ENT) des occurrences totales de marqueurs modaux. Ces résultats montrent que, dans les deux corpus, un groupe très restreint de formes Poly couvre une large quantité d'occurrences. Autrement dit, bien que quantitativement minoritaires dans ModalE, les marqueurs polysémiques semblent occuper une position centrale dans l'expression de la modalité d'un corpus de textes oraux. Nous allons examiner si cela est toujours vrai dans un corpus de textes écrits.

4.2 Cas des corpus écrits

Le premier corpus écrit, nommé TWEETS, est un corpus composé de 11.835 tweets sexistes ou neutres (environ 400.000 tokens) - extrait de (Chiril *et al.*, 2020). Le second corpus, nommé T2K_PEDIA est composé de 502 textes encyclopédiques (environ 400.000 tokens) - extrait de (Battistelli *et al.*, 2022).

	corpus TWEETS			corpus T2K_PEDIA		
	marqueur	*type*	*catégorie(s)*	*marqueur*	*type*	*catégorie(s)*
1	**dire**	Poly	D;E	**pouvoir**	Poly	AL;D;E
2	**pouvoir**	Poly	AL;D;E	grand	Mono	E
3	**bien**	Poly	AP;AX	certain	Poly	AL;E
4	**devoir**	Poly	AL;D;E	**bien**	Poly	AP;AX
5	**vouloir**	Mono	B	permettre	Mono	D

TABLE 2 – TOP5 des marqueurs modaux dans TWEETS et T2K_PEDIA. En surligné : les marqueurs en commun entre ces deux corpus écrits. En gras : les marqueurs en commun avec les corpus oraux.

D'après la ressource ModalE, le corpus TWEETS comporte 505 marqueurs modaux (avec un total de 17.509 occurrences) et le corpus T2K_PEDIA 497 marqueurs modaux (avec un total de 15.724 occurrences).
Nous constatons que le TOP20 des marqueurs les plus fréquents représentent 41% des occurrences des marqueurs modaux dans TWEETS et 44% dans T2K_PEDIA, tandis que les 485 marqueurs restants dans TWEETS et les 477 marqueurs restants dans T2K_PEDIA couvrent respectivement 59% et 56% des occurrences. Bien que les occurrences couvertes par les marqueurs du TOP20 soient plus petites par rapport aux corpus oraux, nous avons décidé de continuer d'appuyer nos analyses des corpus écrits sur les TOP20 afin qu'elles soient comparables avec celles des corpus oraux. Les TOP20 sont disponibles en Annexe B, les TOP5 en Table 2. Détaillons maintenant les profils modaux des deux corpus écrits :

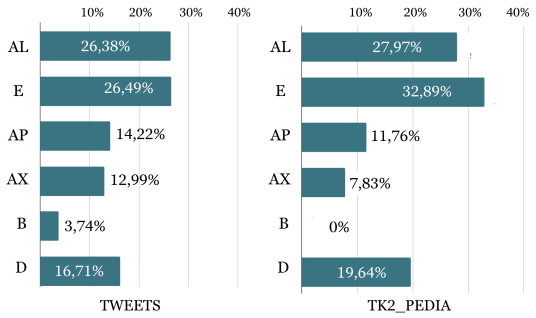

FIGURE 2 – Fréquence normalisée des catégories modales des TOP20 dans les deux corpus écrits

Info_1 [catégories modales exprimées] : parmi les occurrences des marqueurs du TOP20, les modalités les plus fréquentes sont l'épistémique, l'aléthique et la déontique et elles sont majoritairement exprimées par des marqueurs Poly, à l'exception de la modalité épistémique dans T2K_PEDIA (Figure 2). Globalement, nous notons une diminution de la fréquence normalisée de la modalité épistémique par rapport aux corpus oraux, accompagnée, dans TWEETS, par une augmentation des emplois appréciatifs et axiologiques probablement à cause des tweets sexistes mobilisant des adjectifs d'évaluation.
Info_2 [marqueurs modaux employés] : les TOP20 ne partagent que 30% des marqueurs modaux (Annexe B) et deux parmi le TOP5 des plus fréquents (*pouvoir* et *bien*) (Table 2). Les corpus écrits analysés sont donc moins proches entre eux en ce qui concerne les marqueurs modaux employés dans le TOP20, par rapport aux corpus oraux. En revanche, TWEETS partage 50% des marqueurs TOP20 avec les deux corpus d'entretiens ainsi que trois marqueurs parmi les cinq les plus fréquents, alors que T2K_PEDIA ne partage que 20% des marqueurs TOP20. Nous pouvons supposer que le profil modal de TWEETS se rapproche plus de celui des corpus oraux analysés en 4.2. en ce qui concerne les marqueurs employés à cause de son registre plus spontané par rapport à un corpus normé de l'écrit comme l'est T2K_PEDIA.

Info_3 [proportion de marqueurs Poly] : dans les corpus écrits, il y a moins d'occurrences de marqueurs Poly (30% des occurrences dans TWEETS et 25% dans T2K_PEDIA) par rapport aux corpus oraux (40% des occurrences) avec une augmentation des occurrences de marqueurs mixtes [7].

4.3 Ce que partage l'écrit et l'oral dans l'expression de la modalité

Les analyses présentées Section 4.1 et Section 4.2 montrent que l'expression de la modalité, surtout en ce qui concerne les corpus issus de l'oral, dépend essentiellement d'un ensemble restreint de marqueurs prédominants, et qui sont pour la plupart de type Poly. Ces marqueurs dominent l'expression des trois catégories modales les plus fréquentes dans le TOP20 des quatres corpus (à l'exception de T2K_PEDIA en ce qui concerne la modalité épistémique). Ce résultat met en lumière la nécessité de traiter la polysémie de ces marqueurs quand on s'intéresse à déterminer le profil modal d'un texte ou d'un corpus de textes issus de l'oral. Les 4 corpus partagent une liste restreinte de marqueurs en TOP20 (20%, c'est-à-dire 4 marqueurs), dont 2 marqueurs Poly qui apparaissent parmi les 5 marqueurs les plus employés dans tous les corpus : *pouvoir* et *bien*. Ces deux exemples illustrent bien l'hétérogénéité de la catégorie des marqueurs modaux (cf. 3). En effet, il s'agit de deux marqueurs qui peuvent appartenir à des catégories morpho-syntaxiques différentes selon le contexte [8]. Pour (Gosselin, 2010), comme nous l'avons rappelé plus haut, tout marqueur est porteur d'au moins une modalité ; dans cette conception de la modalité qui est celle que nous avons retenue, cela permet alors d'affirmer que *pouvoir* et *bien* sont bien porteurs d'une modalité, quelque soit leur nature morpho-syntaxique. Le travail de désambiguïsation servira ensuite à savoir s'ils marquent toujours la même valeur modale (tout en ayant un statut morpho-syntaxique différent), ou si cette valeur modale change en fonction de la nature morpho-syntaxique et/ou de leur contexte. Par exemple, *bien* est toujours un marqueur Poly appréciatif ou axiologique quand il est un nom, un adjectif ou un adverbe (sa valeur modale en tant que marqueur discursif est en revanche ambiguë, nous avons décidé pour le moment d'exclure ces occurrences du comptage final). Dans cet article, nous allons nous focaliser sur le verbe *pouvoir* qui peut marquer, en tant que coverbe, les modalités aléthique, épistémique et déontique qui sont aussi les 3 modalités les plus fréquentes dans les 4 corpus. On en conclura que la désambiguïsation de ce marqueur est nécessaire pour corriger ou confirmer les proportions des catégories modales employées dans les corpus (cf. Info_1 du profil modal). Nous présentons maintenant le travail effectué sur la désambiguïsation de ce marqueur afin de calculer les valeurs modales qui sont privilégiées dans les 4 corpus et de vérifier alors si la fréquence normalisée des catégories modales exprimées par les marqueurs TOP20 change.

5 Traiter la polysémie de Pouvoir

Afin de désambiguïser les occurrences du marqueur *pouvoir*, nous avons appliqué à nos 4 corpus un modèle obtenu par le fine-tuning de FlauBERT. Le schéma d'annotation utilisé suit également l'approche de (Gosselin, 2010) et rend compte de différentes valeurs sémantiques de pouvoir : éventualité, possibilité matérielle et capacité, sporadicité et permission. Les valeurs de capacité, possibilité matérielle et sporadicité font partie de la catégorie modale aléthique, la valeur d'éventualité fait partie de l'épistémique et la valeur de permission de la déontique. Le modèle (décrit en détail ici

7. Marqueurs Mixtes : 15% dans TWEETS, 19% dans T2K_PEDIA et 8% dans les deux corpus oraux.
8. *Bien* peut être employé en tant que nom, adjectif invariable, adverbe, interjection ou marqueur discursif ; *pouvoir* peut être employé en tant que verbe, coverbe ou nom.

(**?**)) a été entraîné sur une tâche de token classification selon le format IOB avec des données annotées issues du corpus ESLO. Après une première étape d'augmentation des données afin d'équilibrer les classes et une deuxième étape d'augmentation du contexte de la proposition contenant le verbe, le modèle atteint un F1-score (Macro-avg) de 0,92.

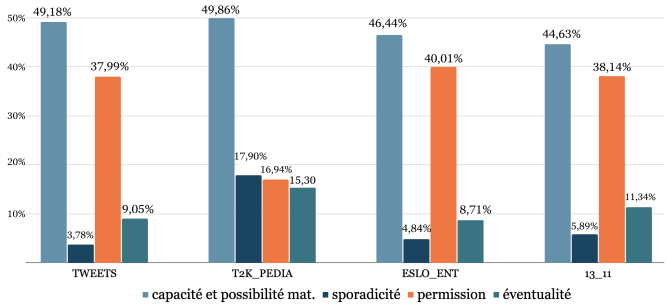

FIGURE 3 – Fréquence normalisée des valeurs de *pouvoir* dans les 4 corpus

En premier lieu, la Figure 3 montre que, dans les deux corpus oraux et dans le corpus TWEETS, la répartition des valeurs modales de pouvoir est similaire avec une prédominance de la valeur de capacité et de celle de possibilité matérielle. Dans T2K_PEDIA les proportions changent : nous observons une diminution de la valeur de permission ainsi qu'une augmentation des occurrences avec valeur de sporadicité qui devient le deuxième emploi le plus fréquent. La valeur sporadique identifie un processus ou un état comme contingent (parfois cet événement se produit et parfois non). Du fait que les textes encyclopédiques sont caractérisés notamment par la présence d'explications, tous les états et les processus qui sont liés à un objet sont recensés comme le montre l'exemple 4.

4. "Une autre difficulté potentielle est que les mailles peuvent glisser à la pointe de l'aiguille quand on pose l'ouvrage" (Wikipedia_Tricot_7)

Ces résultats démontrent une répartition similaire des valeurs des occurrences de *pouvoir* entre les corpus oraux et le corpus TWEETS, ce qui est cohérent avec les similarités des leurs profils modaux (Section 4.2). En revanche, l'analyse qualitative des résultats met en lumière des emplois différents du *pouvoir* de permission dans les corpus. En effet, si, dans TWEETS, nous retrouvons majoritairement des exemples de (non) permission accordée par des lois sociales (cf. ex.5.), dans les corpus oraux, il s'agit plutôt d' emplois d'un pouvoir de politesse (cf. 6.) ou d'un emploi discursif (cf. "on peut dire" dans 7.), typique de l'oral.

5. *À propos de #balancetonporc : "Bientôt on ne pourra plus dire bonjour à une fille."* (TWEETS)

6. *Euh attends j'ai un train de retard tu peux répéter ?* (ENTJEUN_1235)

7. *Enfin j'ai fait essentiellement des mesures on peut dire* (ESLO_ENT_1014)

Dans les corpus 13_11 et ESLO_ENT, la désambiguïsation des occurrences de *pouvoir* n'a pas eu d'impact sur les proportions des catégories modales des TOP20 (Annexe C) alors que nous remarquons des changements en ce qui concerne les corpus écrits. En particulier, la diminution de la fréquence des occurrences de déontique et d'épistémique produit un rapprochement des proportions de l'aléthique et de l'épistémique. Dans TWEETS, c'est désormais l'aléthique qui est la catégorie modale la plus marquée par le TOP20. En ce qui concerne les corpus oraux, nous supposons que la désambiguïsation d'autres marqueurs Poly parmi le TOP20 (p.e. *dire* et *savoir*) pourrait avoir un

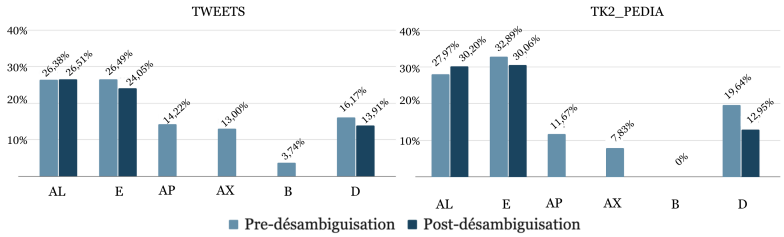

FIGURE 4 – Fréquence normalisée des catégories modales des TOP20 dans les 2 corpus écrits, en pre et post désambiguïsation de pouvoir

impact sur la fréquence des catégories modales marquées.
Il est important de préciser à ce stade que le classifieur employé a été entraîné pour désambiguïser exclusivement les emplois verbaux de pouvoir en tant qu'auxiliaire. En outre, le classificateur fait des erreurs dans la détection et dans la classification de certains emplois de pouvoir qui sont moins représentés dans les données d'entraînement (p.e. les emplois au subjonctif). Une analyse fine des occurrences qui n'ont pas été désambiguisées par le classificateur permettra d'affiner les résultats.

6 Discussion et conclusion

Dans cet article, nous nous sommes focalisées sur les marqueurs modaux lexicaux en démontrant que les marqueurs polysémiques, bien que peu nombreux dans ModalE (5% des marqueurs de la ressource), couvrent une part très importante des occurrences de marqueurs modaux des corpus oraux et écrits - même si plus restreinte pour les corpus écrits [9]. La comparaison entre corpus oraux et écrits a ensuite montré que, d'une part, il y a des différences en ce qui concerne les marqueurs les plus fréquemment employés et la quantité de marqueurs Poly utilisés ; d'autre part, que ces deux types de corpus ont tout de même en commun deux marqueurs dans le TOP5 des marqueurs, l'un d'eux étant le marqueur *pouvoir*. Pour cette raison, nous avons voulu tester l'impact de la désambiguïsation de ce marqueur sur la fréquence normalisée des catégories modales marquées par les TOP20. Les résultats d'un classifieur ont montré une répartition similaire des 4 valeurs modales dans les deux corpus oraux et dans le corpus écrit TWEETS et une répartition différente dans le corpus écrit T2K_PEDIA. Le fait de désambiguïser les occurrences de pouvoir correspond à un changement dans les fréquences normalisées des catégories modales marquées par les TOP20 des corpus écrits. Dans le cadre du travail présenté ici, nous nous sommes focalisées exclusivement sur le *medium* (oral vs écrit), et n'avons pas abordé la question du genre textuel. Dans nos travaux en cours, nous menons un travail de contrôle de l'impact de cette variable sur le changement des profils modaux afin d'affiner nos analyses. Plus généralement, nous souhaitons étendre l'analyse des profils modaux à des corpus de nature différente et souhaitons inclure dans l'analyse la totalité des marqueurs modaux (et non seulement les TOP20) afin d'obtenir des résultats plus complets. Nous développons actuellement deux classifieurs pour deux verbes Poly importants, l'un pour le verbe *devoir*, l'autre pour le verbe *dire*. Ce sont a priori les seuls travaux de ce type menés pour le français. Cet article s'inscrit dans un travail plus ample qui vise à explorer l'expression de la modalité en tant que catégorie sémantique pertinente dans les travaux en psychologie et en psycholinguistique concernant les Troubles du Stress Post-Traumatique (TSPT) (Colli *et al.*, 2024).

9. 42% des occurrences dans 13_11, 41% dans ESLO_ENT, 29% dans TWEETS, 25% dans T2K_PEDIA

Références

BAKER K., BLOODGOOD M., DORR B., FILARDO N. W., LEVIN L. & PIATKO C. (2010). A modality lexicon and its use in automatic tagging. In N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Éds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

BALLY C. (1932). *Linguistique générale et linguistique française*, volume 1. Paris, France : Paris : E. Leroux.

BARBET C. & DE SAUSSURE L. (2012). Présentation : Modalité et évidentialité en français. *Langue française*, **n°173**(1), 3–12. DOI : 10.3917/lf.173.0003.

BATTISTELLI D., ETIENNE A., RAHMAN R., TEISSÈDRE C. & LECORVÉ G. (2022). Une chaîne de traitement pour prédire et appréhender la complexité des textes pour enfants d'un point de vue linguistique (a processing chain to explain the complexity of texts for children from a linguistic and psycho-linguistic point of view). In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 236–246, Avignon, France : ATALA.

BATTISTELLI D. & ÉTIENNE A. (2025). Modale, une ressource lexicale de la modalité au prisme des émotions. *Carnets du Cediscor, numéro Marqueurs modaux, énonciation et argumentation*, **35**.

CHIRIL P., MORICEAU V., BENAMARA F., MARI A., ORIGGI G. & COULOMB-GULLY M. (2020). An annotated corpus for sexism detection in French tweets. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1397–1403, Marseille, France : European Language Resources Association.

COLLI A., BATTISTELLI D. & CHAGNOUX M. (2024). Quel usage des marqueurs modaux dans les discours post-traumatiques ? In *Marqueurs modaux, énonciation et argumentation*, Nantes, France : CerLiCO. HAL : hal-04611485.

GOSSELIN L. (2010). *Les modalités en français : la validation des représentations*, volume 1. Leiden, The Netherlands : Brill. DOI : 10.1163/9789042027572.

NISSIM M., PIETRANDREA P., SANSÒ A. & MAURI C. (2013). Cross-linguistic annotation of modality : a data-driven hierarchical model. In H. BUNT, Éd., *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, p. 7–14, Potsdam, Germany : Association for Computational Linguistics.

PYATKIN V., SADDE S., RUBINSTEIN A., PORTNER P. & TSARFATY R. (2021). The possible, the plausible, and the desirable : Event-based modality detection for language processing. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 953–965, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.77.

RIZVIC-EMINOVIC E. & ŠUKALIĆ (2019). Corpus-based study of the modal verbs in the spoken and academic genres of the corpus of contemporary american english. *Zbornik radova 18*. DOI : 10.51728/issn.2637-1480.2019.17.351.

RUBINSTEIN A., HARNER H., KRAWCZYK E., SIMONSON D., KATZ G. & PORTNER P. (2013). Toward fine-grained annotation of modality in text. In P. PORTNER, A. RUBINSTEIN & G. KATZ,

Éds., *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, p. 38–46, Potsdam, Germany : Association for Computational Linguistics.

RUPPENHOFER J. & REHBEIN I. (2012). Yes we can ! ? annotating English modal verbs. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1538–1545, Istanbul, Turkey : European Language Resources Association (ELRA).

A TOP20 corpus oraux

	corpus 13_11				corpus ESLO_ENT		
	marqueur	*type*	*catégorie(s)*		*marqueur*	*type*	*catégorie(s)*
1	**dire**	Poly	D ;E	1	**bien**	Poly	AP ;AX
2	**savoir**	Poly	AL ;E	2	**dire**	Poly	D ;E
3	**pouvoir**	Poly	AL ;D ;E	3	**savoir**	Poly	AL ;E
4	**penser**	Mono	E	4	**bon**	Poly	AP ;AX
5	**bon**	Poly	AP ;AX	5	**pouvoir**	Poly	AL ;D ;E
6	**bien**	Poly	AP ;AX	6	**falloir**	Poly	AL ;D ;E
7	**vraiment**	Mono	E	7	**vouloir**	Mono	B
8	**vouloir**	Mono	B	8	**petit**	Mono	E
9	**petit**	Mono	E	9	**penser**	Mono	E
10	**personne**	Mono	AL	10	**croire**	Mono	E
11	**arriver**	Mono	AL	11	toujours	Mono	AL
12	**falloir**	Poly	AL ;D ;E	12	**peut-être**	Mono	E
13	**croire**	Mono	E	13	travail	Mixte	AL ;AP
14	**devoir**	Poly	AL ;D ;E	14	aimer	Mono	AP
15	**venir**	Mono	AL	15	**devoir**	Poly	AL ;D ;E
16	**essayer**	Mixte	B ;E	16	travailler	Mixte	AL ;AP
17	connaître	Mono	E	17	**arriver**	Mono	AL
18	**peut-être**	Mono	E	18	**vraiment**	Mono	E
19	rien	Mono	AL	19	**venir**	Mono	AL
20	juste	Poly	AX ;E	20	vrai	Mono	E

TABLE 3 – TOP20 marqueurs modaux dans 13_11 et ESLO_ENT. En gras : les marqueurs en commun entre les deux corpus.

B TOP20 corpus écrits

	corpus TWEETS				corpus T2K_PEDIA		
	marqueur	*type*	*catégorie(s)*		*marqueur*	*type*	*catégorie(s)*
1	**dire**	Poly	D ;E	1	**pouvoir**	Poly	AL ;D ;E
2	**pouvoir**	Poly	AL ;D ;E	2	grand	Mono	E
3	**bien**	Poly	AP ;AX	3	certain	Poly	AL ;E
4	**devoir**	Poly	AL ;D ;E	4	**bien**	Poly	AP ;AX
5	**vouloir**	Mono	B	5	permettre	Mono	D
6	viol	Mono	AX	6	**devoir**	Poly	AL ;D ;E
7	travail	Mixte	AL ;AP	7	selon	Mono	E
8	**falloir**	Poly	AL ;D ;E	8	petit	Mono	E
9	**bon**	Poly	AP ;AX	9	guerre	Mixte	AL ;AX
10	**savoir**	Poly	AL ;E	10	travail	Mixte	AL ;AP
11	rien	Mono	AL	11	connaître	Mono	E
12	harcèlement	Mono	AX	12	souvent	Mono	AL
13	**venir**	Mono	AL	13	considérer	Mono	E
14	beau	Mono	AP	14	long	Mono	E
15	droit	Mono	D	15	**dire**	Poly	D ;E
16	personne	Mono	AL	16	mort	Mixte	AL ;AP
17	grand	Mono	E	17	personne	Mono	AL
18	**penser**	Mono	E	18	toujours	Mono	AL
19	aimer	Mono	AP	19	parfois	Mono	AL
20	toujours	Mono	AL	20	blanc	Mono	AL

TABLE 4 – TOP20 marqueurs modaux dans TWEETS et T2K_PEDIA. En gras : les marqueurs en commun avec les corpus oraux. En surligné : les marqueurs en commun entre TWEETS et T2K_PEDIA.

C Pre/Post désambiguisation corpus oraux

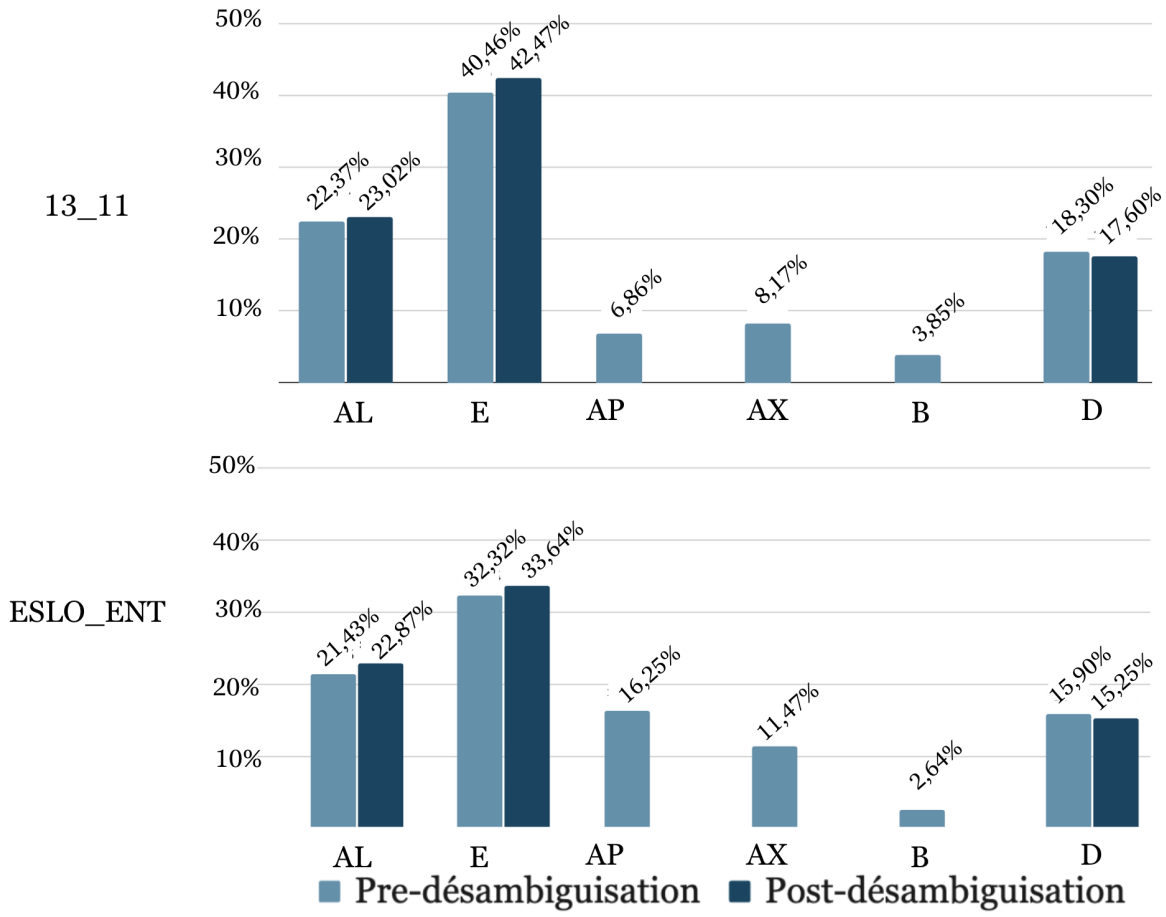

FIGURE 5 – La fréquence normalisée des catégories modales des TOP20 marqueurs dans le corpus 13_11 et ESLO_ENT pre et post désambiguïsation de pouvoirs