LA LINGUISTIQUE PRÉDICTIVE : CONCEPTS, MÉTHODES ET APPLICATIONS

M. GUIDERE. Professeur des Universités. Directeur de recherches à l'INSERM

Institut National de la Santé et de la Recherche Médicale, France. mathieu.guidere@inserm.fr

Résumé : Cet article propose une synthèse intégrée de la linguistique prédictive comme cadre théorique et méthodologique pour l'étude du langage naturel et le développement des grands modèles de langage. Il montre que les phénomènes linguistiques, du niveau sublexical au discours étendu, peuvent être décrits comme des processus d'anticipation fondés sur des régularités distributionnelles.

L'analyse prédictive se déploie à travers les différents niveaux de structuration linguistique. Au niveau morphologique, elle met en évidence la capacité à anticiper la structure interne des mots à partir de régularités propres à des langues variées, qu'elles soient concaténatives, agglutinantes ou non concaténatives. Au niveau syntaxique, elle montre que l'ordre des mots et les relations grammaticales peuvent être modélisés comme des contraintes probabilistes, permettant de rendre compte de la grammaticalité comme phénomène graduel. Sur le plan sémantique, le sens est conçu comme une distribution dynamique, dépendante du contexte et des cooccurrences, tandis que la pragmatique introduit une dimension inférentielle liée aux intentions et aux normes interactionnelles.

L'article accorde une attention particulière à la structuration informationnelle et discursive. La distinction entre thème et rhème est reformulée en termes de prévisibilité, et le discours est analysé comme une succession de segments contraints par des relations de cohérence et des schémas de genre. Une méthodologie spécifique est proposée, fondée sur l'exploitation de corpus multilingues, l'annotation de relations discursives et l'entraînement de modèles capables de rendre compte des régularités à grande échelle.

Les applications de cette approche sont ensuite examinées dans divers domaines. Elles incluent la génération et la traduction automatiques, l'analyse des émotions, l'étude des pratiques culturelles, ainsi que les usages en santé mentale, où l'analyse prédictive du discours permet d'identifier des configurations linguistiques associées à des états de souffrance.

Mots clés : Linguistique prédictive, modèles de langage, cognition, morphologie, syntaxe, sémantique, pragmatique, discours, intelligence artificielle, multilinguisme, culture, émotions, santé mentale.

Abstract: This article proposes an integrated synthesis of predictive linguistics as a theoretical and methodological framework for the study of natural language and the development of large language models. It shows that linguistic phenomena, from the sublexical level to extended discourse, can be described as processes of anticipation grounded in distributional regularities.

Predictive analysis unfolds across the different levels of linguistic structure. At the morphological level, it highlights the ability to anticipate the internal structure of words based on regularities specific to diverse languages, whether concatenative, agglutinative, or non-concatenative. At the syntactic level, it demonstrates that word order and grammatical relations can be modeled as probabilistic constraints, accounting for grammaticality as a gradient phenomenon. At the semantic level, meaning is conceived as a dynamic distribution dependent on context and co-occurrence patterns, while pragmatics introduces an inferential dimension linked to communicative intentions and interactional norms.

The article pays particular attention to informational and discourse structuring. The distinction between theme and rheme is reformulated in terms of predictability, and discourse is analyzed as a sequence of segments constrained by coherence relations and genre-specific patterns. A specific methodology is proposed, based on the use of multilingual corpora, the annotation of discourse relations, and the training of models capable of capturing large-scale regularities.

The applications of this approach are then examined across various domains. These include text generation and machine translation, emotion analysis, the study of cultural practices, as well as applications in mental health, where predictive discourse analysis enables the identification of linguistic configurations associated with states of distress.

Keywords: Predictive linguistics, language models, cognition, morphology, syntax, semantics, pragmatics, discourse, artificial intelligence, multilingualism, culture, emotions, mental health.

Introduction

La linguistique prédictive constitue aujourd'hui un cadre conceptuel central pour comprendre le fonctionnement et le développement des grands modèles de langage. Elle repose sur l'hypothèse selon laquelle la production et la compréhension du langage humain peuvent être décrites comme des processus d'anticipation probabiliste, où les unités linguistiques sont générées ou interprétées en fonction de leur probabilité conditionnelle dans un contexte donné.

Dans cette perspective, le langage n'est plus envisagé comme un système purement symbolique régi par des règles abstraites, mais comme une distribution statistique sur des séquences d'unités, qu'il s'agisse de phonèmes, de morphèmes ou de mots.

Cette conception trouve un ancrage empirique dans les travaux en psycholinguistique et en neurolinguistique montrant que les locuteurs humains anticipent activement les éléments à venir lors du traitement du langage, en mobilisant des connaissances syntaxiques, sémantiques et pragmatiques intégrées.

Les grands modèles de langage (LLMs), fondés sur des architectures neuronales profondes de type transformeur, opérationnalisent directement cette hypothèse prédictive. Leur objectif d'apprentissage consiste à estimer la probabilité d'apparition d'un token donné conditionnellement à une séquence préalable, ce qui correspond formellement à une modélisation du langage comme processus stochastique séquentiel. Ainsi, la tâche d'entraînement dite de prédiction du prochain mot constitue une instanciation computationnelle de la linguistique prédictive.

Ce cadre présente plusieurs implications théoriques. D'une part, il tend à estomper la frontière entre compétence linguistique et performance, en intégrant les variations d'usage dans le modèle lui-même. D'autre part, il redéfinit la notion de grammaticalité en termes de probabilité plutôt que de conformité stricte à des règles, ce qui permet de rendre compte de phénomènes gradiels et contextuels souvent difficiles à modéliser dans des approches formelles classiques.

Sur le plan méthodologique, la linguistique prédictive offre un principe unificateur pour l'apprentissage non supervisé à grande échelle. En exploitant des corpus massifs, les modèles apprennent des régularités multi-niveaux sans annotation explicite, ce qui suggère que des propriétés linguistiques complexes peuvent émerger de la simple optimisation d'un objectif prédictif.

Cette émergence pose toutefois des questions quant à la nature des représentations internes, notamment en ce qui concerne leur interprétabilité et leur alignement avec des catégories linguistiques traditionnelles.

Enfin, la linguistique prédictive soulève des enjeux épistémologiques. Si elle permet de modéliser efficacement le comportement linguistique observable, elle ne garantit pas une compréhension des mécanismes cognitifs sous-jacents. En effet, la convergence entre performances des modèles et comportements humains ne doit pas être interprétée comme une équivalence de processus, mais plutôt comme une analogie fonctionnelle. Dès lors, la linguistique prédictive apparaît moins comme une théorie exhaustive du langage que comme un paradigme opératoire, particulièrement fécond pour la recherche contemporaine sur les modèles de langage à grande échelle.

Le cerveau comme machine à prédictions

La convergence entre linguistique prédictive et sciences cognitives s'inscrit dans un cadre théorique plus large, souvent désigné sous le terme de traitement prédictif, selon lequel le cerveau humain fonctionne comme un système d'inférence bayésienne hiérarchique. Dans cette perspective, la cognition est fondamentalement orientée vers la minimisation de l'erreur de prédiction entre des attentes internes et des entrées sensorielles effectives.

Appliquée au langage, cette approche conduit à considérer que la compréhension linguistique repose sur une génération continue d'hypothèses probabilistes concernant les unités à venir, à différents niveaux de représentation.

Les données empiriques issues de la neurocognition du langage appuient cette hypothèse. Des marqueurs électrophysiologiques tels que la composante N400 ont été interprétés comme reflétant des coûts de traitement liés à des violations d'attentes sémantiques, suggérant que le cerveau anticipe activement les propriétés du mot suivant en contexte. De plus, des études en imagerie cérébrale ont mis en évidence l'implication de réseaux distribués dans la mise à jour dynamique des prédictions linguistiques, impliquant notamment les régions temporales et frontales.

Dans ce cadre, le traitement linguistique peut être décrit comme un processus de mise à jour itérative de modèles internes génératifs, où les représentations sont ajustées en fonction des écarts entre prédictions et entrées effectives.

Cette dynamique implique une hiérarchie de niveaux, allant des caractéristiques acoustiques aux structures syntaxiques et aux inférences pragmatiques, chaque niveau contraignant les autres dans un système d'interactions bidirectionnelles. La prédiction ne se limite donc pas à l'anticipation lexicale, mais englobe l'ensemble des dimensions du langage.

La comparaison avec les grands modèles de langage met en évidence des analogies fonctionnelles, mais également des divergences structurelles. Comme ces modèles, le cerveau semble exploiter des régularités statistiques pour générer des attentes contextuelles. Toutefois, contrairement aux architectures artificielles, le système cognitif humain intègre des contraintes issues de l'expérience sensorimotrice, de l'intentionnalité et de l'interaction sociale, ce qui confère aux prédictions une dimension incarnée et située.

Sur le plan théorique, cette approche contribue à redéfinir la notion de compétence linguistique en l'inscrivant dans une dynamique adaptative continue. La connaissance linguistique ne se réduit plus à un ensemble de règles stockées, mais correspond à un ensemble de distributions probabilistes constamment réajustées en fonction de l'expérience. Cette conception permet de rendre compte de la flexibilité et de la robustesse du traitement du langage, notamment face à l'ambiguïté et à la variabilité.

Enfin, l'hypothèse du cerveau comme machine à prédictions soulève des questions quant à la nature des représentations mentales et à leur niveau d'abstraction. Si les modèles prédictifs rendent compte de nombreux phénomènes comportementaux et neurophysiologiques, ils laissent ouverte la question de savoir dans quelle mesure ces prédictions reposent sur des structures symboliques explicites ou sur des représentations distribuées.

La morphologie prédictive

Appliquée à la morphologie, la linguistique prédictive conduit à envisager le mot non comme une unité indivisible, mais comme une séquence structurée d'indices formels dont l'agencement permet d'anticiper des propriétés grammaticales et sémantiques.

Dans cette perspective, l'identification d'un radical, d'un préfixe, d'un suffixe ou d'un marqueur flexionnel relève d'un traitement probabiliste fondé sur la régularité des cooccurrences observées dans les données.

L'analyse morphologique devient ainsi une opération d'inférence, par laquelle le système estime la structure interne la plus plausible d'une forme donnée à partir de patrons récurrents.

Cette approche est particulièrement productive dans les langues à morphologie concaténative. Par exemple en français, la reconnaissance du radical « chant » dans « chanter, chanteur, chanson, rechanter… » permet d'anticiper une parenté lexicale, tandis que les suffixes « -eur », « -ion » ou les désinences verbales orientent l'interprétation morphosyntaxique.

De même, dans « défaire » ou « impossible », les préfixes « dé » et « im » activent des attentes relatives à l'inversion ou à la négation.

En anglais, des formes comme « unhappiness », « replay » ou « washable » illustrent le même principe : la détection de « un », « re » et « able » permet d'inférer respectivement la négation, l'itération et la possibilité. Le traitement prédictif repose ici sur la capacité à segmenter la chaîne et à attribuer à chaque segment une fonction probable.

En espagnol et en italien, la richesse flexionnelle renforce encore la portée prédictive de la morphologie. En espagnol, une forme telle que « hablábamos » permet d'anticiper, avant même l'analyse complète, un radical verbal « habl » associé au passé imperfectif, à la première personne du pluriel et à un verbe du premier groupe.

De même, « nacionalización » se prête à une décomposition en « nacional », « iza », « ción », où chaque segment contribue à la construction compositionnelle du sens.

En italien, « impossibile », « rileggere » ou « velocemente » montrent comment les affixes « im », « ri » et « mente » orientent respectivement vers la négation, la répétition et l'adverbialisation.

Dans ces langues, la prévisibilité des paradigmes flexionnels et dérivationnels fournit un socle robuste à l'inférence morphologique.

Les langues non indoeuropéennes montrent que ce modèle prédictif demeure opératoire, mais qu'il doit s'adapter à d'autres types de structuration. Par exemple, en japonais, la morphologie agglutinante offre des séquences particulièrement transparentes pour l'anticipation. Une forme comme 食べさせられました, tabesaseraremashita, peut être analysée comme une succession ordonnée de morphèmes exprimant le radical verbal, le causatif, le passif ou potentiel selon l'analyse adoptée, puis la politesse et le passé.

Le système peut ainsi exploiter la régularité de l'ordre morphémique pour prédire les unités suivantes à partir des premières.

De même, dans 読みたい, yomitai, l'apparition de « tai » après le radical verbal permet d'anticiper une

valeur désidérative.

La morphologie japonaise illustre bien une prédiction séquentielle à forte contrainte positionnelle.

En chinois, où la morphologie flexionnelle est limitée, la prédiction porte moins sur des affixes que sur la composition lexicale et la récurrence de morphèmes liés. Des formes comme 电脑, diannao, ordinateur, littéralement « électricité plus cerveau », ou 国际化, guojihua, internationalisation, montrent que l'interprétation d'un mot peut reposer sur l'identification de morphèmes récurrents tels que 化, hua, marqueur de transformation.

De même, dans 现代化, xiandaihua, modernisation, la présence de « hua » permet d'anticiper un procès de changement d'état ou de mise en conformité.

La linguistique prédictive appliquée au chinois doit donc intégrer une morphologie davantage compositionnelle que flexionnelle, où la fréquence des associations entre caractères joue un rôle central.

Le cas de l'arabe est particulièrement instructif, car il met en jeu une morphologie non concaténative fondée sur l'interaction entre racines consonantiques et schèmes vocaliques. Une séquence comme « k t b » active un champ lexical lié à l'écriture, que l'on retrouve dans kataba, il a écrit, kitaab, livre, kâtib, écrivain, ou maktab, bureau. Ici, la prédiction morphologique ne repose pas seulement sur la juxtaposition linéaire d'affixes, mais sur la reconnaissance d'un patron abstrait associant racine et gabarit.

L'identification d'un schème comme maCCaC ou CaaCiC permet d'anticiper une catégorie lexicale ou une fonction dérivationnelle.

Ce type de morphologie montre que la prédiction peut porter sur des dépendances discontinues et sur des structures abstraites qui excèdent la simple segmentation linéaire.

Du point de vue des grands modèles de langage, l'analyse prédictive de la morphologie présente un double intérêt. D'une part, elle améliore la représentation des mots rares, complexes ou inconnus en exploitant leur structure interne. D'autre part, elle permet une meilleure généralisation translexicale, puisqu'un modèle capable d'identifier des régularités morphémiques peut transférer des connaissances d'une forme à une autre apparentée. Cette propriété est particulièrement utile dans les langues à forte productivité dérivationnelle ou flexionnelle, où le nombre de formes attestées excède largement ce qu'un apprentissage purement lexical pourrait mémoriser.

La syntaxe prédictive

La syntaxe prédictive envisage la structuration phrastique comme un processus d'anticipation hiérarchisée, dans lequel chaque élément lexical contraint probabilistiquement l'apparition et la position des éléments suivants.

Plutôt que de postuler un ensemble de règles discrètes appliquées de manière séquentielle, cette approche décrit la formation des phrases comme une navigation dans un espace de configurations possibles, pondérées par leur vraisemblance contextuelle.

La grammaticalité y apparaît comme un gradient de conformité aux attentes générées par les distributions syntaxiques internalisées.

En français, l'ordre canonique « sujet, verbe, objet » fournit un cadre prédictif fort, mais modulé par des phénomènes d'accord et de dépendance à distance. Dans une séquence comme « les enfants mangent », la présence du syntagme nominal pluriel « les enfants » induit une forte attente pour une forme verbale accordée au pluriel, ce qui rend « mangent » nettement plus probable que « mange ».

Si l'on prolonge par une expansion, « les enfants mangent une pomme », l'introduction du verbe transitif active l'attente d'un complément d'objet direct. Une déviation telle que « les enfants mangent une » serait perçue comme incomplète, non en vertu d'une règle violée abstraitement, mais parce qu'elle ne satisfait pas les contraintes prédictives associées au verbe « manger ».

Les phénomènes d'inversion interrogative, comme dans « mange-t-il une pomme », illustrent également une reconfiguration des attentes, où la position du verbe en tête signale une structure interrogative et modifie la distribution attendue des clitiques.

En anglais, la relative rigidité de l'ordre des mots renforce la dimension prédictive de la syntaxe. Une séquence comme « the cat chased » implique fortement l'apparition d'un objet, comme dans « the cat chased the mouse ». L'omission de cet objet dans un contexte non intransitif crée une tension prédictive.

De même, dans « a very interesting book », la présence de « very » anticipe un adjectif gradable, ce qui rend une continuation comme « a very book » improbable.

Les constructions interrogatives, telles que « what did you see », montrent comment l'auxiliaire « did » signale une structure spécifique qui reconfigure les attentes quant à l'ordre sujet verbe.

La syntaxe prédictive en anglais s'appuie ainsi sur des patrons de surface relativement stables, facilitant l'inférence séquentielle.

Les langues romanes comme l'espagnol et l'italien introduisent une flexibilité relative de l'ordre des constituants, compensée par des indices morphologiques riches.

En espagnol, une phrase comme « comieron los niños la manzana », où le verbe précède le sujet, reste interprétable grâce à l'accord verbal « comieron », troisième personne du pluriel, qui permet d'anticiper un sujet pluriel, ici « los niños ». La présence d'un objet défini « la manzana » est également facilitée par la structure transitive du verbe.

Dans une autre configuration, « los niños comieron », l'absence d'objet peut être interprétée comme une ellipse pragmatique, car le verbe permet cette omission.

En italien, « ha visto Maria » peut être interprété comme « il ou elle a vu Maria », mais aussi, dans certains contextes, comme « Maria a vu quelqu'un » si l'ordre est marqué et si l'intonation le suggère. La prédiction syntaxique doit donc intégrer des indices multiples, incluant l'accord, la position et le contexte discursif.

En japonais, la syntaxe prédictive repose sur une organisation fortement marquée par les particules et une structure de type « sujet, objet, verbe ».

Dans une phrase comme 太郎がリンゴを食べた, Tarō ga ringo o tabeta, la particule « ga » signale le sujet, tandis que « o » marque l'objet direct. Dès l'apparition de « ringo o », le système anticipe un verbe transitif en position finale.

L'ordre des constituants peut être relativement flexible, mais les particules maintiennent la clarté des relations syntaxiques.

Dans une séquence comme 太郎がリンゴを, l'absence du verbe final est perçue comme une suspension, car la particule « o » crée une attente forte pour un prédicat.

De même, l'introduction d'une subordonnée avec と, to, comme dans 行くと思う, iku to omou, « je pense qu'il va partir », permet d'anticiper une structure enchâssée où la proposition précédente fonctionne comme complément du verbe de pensée.

Le chinois présente une syntaxe où l'ordre des mots joue un rôle central, en l'absence de marquage flexionnel. Une phrase comme 他吃苹果, tā chī píngguǒ, il mange une pomme, repose sur un ordre « sujet, verbe, objet » relativement strict. L'introduction d'un adverbe aspectuel, comme 在 dans 他在吃苹果, il est en train de manger une pomme, modifie les attentes en signalant une progression en cours, ce qui rend certaines continuations plus probables que d'autres.

Les constructions à thème, comme 这本书我看过, « ce livre, je l'ai lu », illustrent une réorganisation des attentes, où le syntagme initial définit le domaine de pertinence de la prédication suivante.

La prédiction syntaxique en chinois s'appuie ainsi sur des patrons positionnels et des marqueurs fonctionnels non flexionnels.

L'arabe, enfin, combine une richesse morphologique avec une certaine variabilité de l'ordre des mots, notamment entre les structures « verbe, sujet, objet » et « sujet, verbe, objet ».

Dans une phrase comme كتب الطالب الرسالة, kataba al ṭālibu ar-risālata, « a écrit l'étudiant la lettre », l'ordre « verbe, sujet, objet » est canonique dans de nombreux contextes. La forme verbale kataba, troisième personne du singulier masculin, permet d'anticiper un sujet compatible en genre et en nombre.

Si l'on rencontre une forme comme كتبوا, katabū, ils ont écrit, l'attente se porte sur un sujet pluriel masculin ou mixte.

La présence de marques casuelles, lorsque réalisées, et de l'accord verbal contribue à la désambiguïsation des rôles syntaxiques, même lorsque l'ordre des constituants varie.

Cette capacité prédictive repose sur l'intériorisation de contraintes statistiques fines, qui capturent à la fois des dépendances locales et des relations à longue distance, et qui rendent compte du caractère à la fois structuré et flexible de la syntaxe naturelle.

La sémantique prédictive

La sémantique prédictive conçoit l'interprétation du sens comme un processus d'inférence probabiliste guidé par le contexte, dans lequel les unités lexicales activent des ensembles de significations possibles pondérées par leur compatibilité avec les attentes en cours.

Le sens n'est plus envisagé comme une propriété fixe attachée aux mots, mais comme une distribution dynamique de valeurs sémantiques, continuellement ajustée à mesure que la séquence linguistique se déploie.

Cette approche permet d'intégrer, dans un même cadre, la polysémie, l'ambiguïté contextuelle et la composition du sens à l'échelle de la phrase.

Au niveau lexical, la prédiction sémantique repose sur la cooccurrence et la similarité distributionnelle. Par

exemple en français, le mot « banque » active des interprétations distinctes selon le contexte immédiat. Dans « elle travaille à la banque », l'environnement professionnel oriente vers l'institution financière, tandis que dans « ils travaillent sur une banque de données », la présence de « données » active une interprétation informatique.

Le traitement prédictif consiste ici à ajuster la probabilité des sens concurrents en fonction des indices contextuels.

En anglais, un phénomène analogue apparaît avec « bank » ou « bat », où la désambiguïsation dépend de cooccurrences telles que « money », « river » ou « baseball ». Dans "she went to the bank to withdraw cash", la présence de "withdraw" et "cash" réduit fortement l'incertitude sémantique.

Dans les langues romanes comme l'espagnol et l'italien, la morphologie dérivationnelle contribue à orienter la prédiction du sens.

En espagnol, des formes comme « feliz », « felicidad » et « felizmente » partagent un noyau sémantique commun lié à l'idée de bonheur, tandis que les suffixes modulent la catégorie grammaticale et la fonction discursive. Dans « una decisión importante », le mot « importante » anticipe une évaluation qualitative, mais sa portée exacte dépend du nom qu'il modifie.

En italien, « leggere » peut signifier « lire » ou « être léger » selon le contexte morphosyntaxique, comme dans « leggere un libro » ou « una borsa leggera ». La prédiction sémantique implique donc une interaction étroite entre indices lexicaux et structurels.

Les langues à forte dépendance contextuelle, comme le japonais et le chinois, illustrent de manière particulièrement nette la nature inférentielle du sens.

En japonais, un énoncé comme 彼は橋を渡った, kare wa hashi o watatta, peut contenir une ambiguïté phonétique entre 橋, pont, et 箸, baguettes, bien que l'écriture la résolve.

Dans un contexte oral, la prédiction sémantique s'appuie sur la plausibilité situationnelle, rendant l'interprétation pont beaucoup plus probable avec le verbe « traverser ».

De même, dans 彼は本を読んでいる, kare wa hon o yonde iru, il est en train de lire un livre, la combinaison de 本, livre, et 読む, lire, constitue une collocation fortement attendue, réduisant l'incertitude interprétative.

En chinois, la composition morphémique et l'absence de flexion accentuent le rôle du contexte dans la construction du sens. Une séquence comme 开车, kāi chē, conduire une voiture, associe « ouvrir » et « véhicule », mais l'interprétation idiomatique est prédite par la fréquence d'usage. Dans 他开了一个公司, tā kāi le yī gè gōngsī, il a fondé une entreprise, le verbe 开 prend un sens abstrait de création ou de gestion, activé par le complément 公司, entreprise. La particule aspectuelle 了 contribue également à situer l'événement dans le temps, influençant l'interprétation globale de l'énoncé. La sémantique prédictive en chinois repose ainsi sur des associations lexicales et des schémas d'usage stabilisés.

L'arabe offre un autre type de régularité, où la structure morphologique interagit avec la prédiction sémantique. La racine « k t b », liée à l'écriture, génère des formes comme كتاب, kitāb, livre, مكتب, maktab, bureau, ou كاتب, kātib, écrivain. La reconnaissance de cette racine active un champ sémantique cohérent, permettant d'anticiper des significations liées à l'acte d'écrire ou à ses produits.

Dans une phrase comme كتب الطالب الرسالة, kataba al ṭālibu al risāla, l'étudiant a écrit la lettre, la combinaison du verbe « écrire » et du nom « lettre » constitue une association fortement prédite. La sémantique prédictive doit ici intégrer des correspondances entre structures morphologiques abstraites et domaines conceptuels.

Au niveau phrastique, la composition du sens repose sur l'intégration progressive des contraintes lexicales, syntaxiques et pragmatiques. Une phrase comme « le chat a mangé la souris » active une structure événementielle prototypique où un agent animé agit sur un patient, ce qui rend l'interprétation directe et peu ambiguë.

En revanche, dans « le chat a mangé la table », la compatibilité sémantique est faible, ce qui entraîne une réévaluation interprétative, par exemple vers une lecture métaphorique ou humoristique.

En anglais, the « idea devoured him » illustre une inversion des rôles sémantiques attendus, où un concept abstrait occupe la position d'agent, ce qui reste interprétable grâce à des extensions métaphoriques.

Dans des langues comme le japonais, la structure informationnelle et le contexte discursif jouent un rôle déterminant. Une phrase comme 雨が降っている, ame ga futte iru, il pleut, peut être interprétée de manière neutre, mais dans un contexte donné, elle peut impliquer des inférences pragmatiques, comme la nécessité de prendre un parapluie.

En chinois, une séquence comme 他来了, tā lái le, il est venu, peut également véhiculer des implications contextuelles, telles qu'un changement d'état pertinent pour la situation communicative. La

prédiction sémantique ne se limite donc pas au contenu propositionnel, mais inclut les effets pragmatiques attendus.

Du point de vue des modèles de langage, la sémantique prédictive se traduit par la capacité à associer des représentations vectorielles aux unités linguistiques, de manière à capter des régularités de sens à partir de leur distribution contextuelle. Ces modèles apprennent à ajuster en continu les probabilités associées aux interprétations possibles, en fonction des contextes observés.

Cette dynamique permet de rendre compte de la polysémie, de la métaphore et de la composition du sens sans recourir à des définitions explicites, mais elle soulève également des questions quant à la nature des représentations obtenues et à leur correspondance avec des catégories sémantiques traditionnelles.

La pragmatique prédictive

La pragmatique prédictive étend le cadre probabiliste à l'usage effectif du langage en contexte, en considérant que l'interprétation d'un énoncé repose sur l'anticipation des intentions communicatives et des effets attendus dans une situation donnée. Le sens pragmatique n'est pas directement codé dans les formes linguistiques, mais inféré à partir d'indices contextuels, sociaux et interactionnels. Le locuteur et l'interlocuteur sont ainsi modélisés comme des agents qui génèrent et ajustent des hypothèses sur les états mentaux, les buts et les connaissances partagées, selon une logique d'inférence graduelle.

En français, des énoncés apparemment simples illustrent cette dynamique. Une phrase comme « il fait froid ici », produite dans une pièce dont la fenêtre est ouverte, peut être interprétée non comme une simple description météorologique, mais comme une requête indirecte visant à inciter l'interlocuteur à fermer la fenêtre. La probabilité de cette interprétation dépend du contexte physique, de la relation entre les participants et des normes interactionnelles.

De même, « tu pourrais fermer la porte » présente une structure interrogative, mais active fortement une lecture directive, car la forme modale « pourrait » est conventionnellement associée à une demande polie. La pragmatique prédictive consiste ici à évaluer la plausibilité relative de différentes intentions sous-jacentes.

En anglais, des phénomènes comparables apparaissent dans des formulations telles que « can you pass the salt », qui, malgré sa forme interrogative, est systématiquement interprétée comme une requête. La présence de « can you » n'active pas réellement une question sur la capacité physique, mais une convention pragmatique stabilisée.

Dans un autre registre, l'énoncé « it's getting late » peut fonctionner comme un signal implicite de clôture d'interaction, notamment dans un contexte social où l'un des participants souhaite mettre fin à la rencontre. La prédiction pragmatique s'appuie ici sur des scripts interactionnels et des attentes culturelles.

Les langues romanes comme l'espagnol et l'italien offrent des exemples où la morphologie et l'intonation interagissent avec la pragmatique.

En espagnol, « quieres cerrar la ventana » peut être interprété comme une question authentique ou comme une suggestion atténuée selon le contexte et la prosodie. De même, « claro » peut signifier « bien sûr », mais aussi fonctionner comme marqueur discursif exprimant l'évidence ou parfois une légère impatience.

En italien, « magari » peut exprimer un souhait positif ou une réponse ironique selon la situation. Dans ces cas, la prédiction pragmatique nécessite l'intégration d'indices subtils, souvent extralinguistiques.

En japonais, la pragmatique prédictive est fortement structurée par les systèmes de politesse et par l'implicite. Une phrase comme ちょっと難しいですね, chotto muzukashii desu ne, littéralement « c'est un peu difficile », peut fonctionner comme un refus indirect, par exemple dans une négociation ou une invitation. L'interprétation dépend de la convention culturelle selon laquelle l'expression directe du refus est souvent évitée.

De même, l'omission fréquente du sujet oblige l'interlocuteur à inférer les référents à partir du contexte partagé. Dans 行きますか, ikimasu ka, « allez-vous y aller », l'absence de sujet explicite n'entrave pas la compréhension, car la situation interactionnelle permet de prédire l'agent pertinent.

Le chinois met également en évidence le rôle central du contexte dans l'interprétation pragmatique. Une réponse comme 可以, kěyǐ, peut signifier « d'accord, c'est possible » ou même une autorisation, selon la situation. Dans 你吃了吗, nǐ chī le ma, « as-tu mangé », la question peut fonctionner comme une formule de salutation dans certains contextes, sans attente réelle d'information. La particule finale 吧, ba, comme dans 我们走吧, wǒmen zǒu ba, partons, introduit une nuance de suggestion ou d'invitation, modulant la force illocutoire de l'énoncé. La prédiction pragmatique repose ici sur des marqueurs discursifs et des routines interactionnelles.

L'arabe illustre quant à lui des usages pragmatiques où la formule linguistique excède son contenu littéral.

Une expression comme إن شاء الله, in shā' Allāh, si Dieu le veut, peut exprimer une intention sincère, mais aussi, dans certains contextes, une réserve ou une absence d'engagement ferme. L'interprétation dépend des attentes culturelles et du contexte conversationnel.

De même, la répétition ou l'insistance dans certaines formules peut signaler la politesse, l'urgence ou l'ironie, selon les normes sociales en vigueur.

Au niveau interactionnel, la pragmatique prédictive implique une modélisation dynamique du contexte commun, incluant les connaissances partagées, les croyances mutuelles et les objectifs communicatifs. Une réponse minimale comme d'accord peut, selon le moment et l'intonation, signaler l'acceptation, la résignation ou même une forme de désaccord implicite. L'interlocuteur doit donc anticiper non seulement le contenu propositionnel, mais aussi la valeur illocutoire et les effets perlocutoires attendus.

Dans le cadre des grands modèles de langage, cette dimension se traduit par la capacité à ajuster les productions en fonction de contextes discursifs élargis, incluant des tours de parole antérieurs et des indices situationnels implicites. Les modèles apprennent des régularités pragmatiques à partir de corpus, ce qui leur permet de générer des réponses appropriées dans de nombreux contextes simulés. Toutefois, l'absence d'ancrage dans une expérience vécue limite leur accès à certaines dimensions de l'intentionnalité et de la norme sociale, qui restent partiellement implicites dans les données.

Analyse prédictive de l'énonciation

L'analyse prédictive de l'énonciation s'inscrit dans une approche informationnelle du langage où la structuration des énoncés est interprétée comme une gestion dynamique des attentes relatives à ce qui est déjà établi dans le discours et à ce qui est introduit comme apport informatif.

La distinction entre thème et rhème, souvent décrite comme opposition entre information connue et information nouvelle, peut être reformulée en termes probabilistes : le thème correspond à des éléments à forte prévisibilité contextuelle, tandis que le rhème constitue une zone de réduction d'incertitude, apportant des informations moins attendues.

Dans cette perspective, la production et la compréhension des énoncés impliquent une anticipation continue de la structure informationnelle.

En français, une phrase comme ce livre, je l'ai déjà lu illustre une organisation où ce livre est posé comme thème, activant une référence présupposée ou accessible, tandis que je l'ai déjà lu constitue le rhème, apportant une information sur cet objet. La dislocation à gauche signale explicitement cette structuration, permettant à l'interlocuteur de calibrer ses attentes dès le début de l'énoncé.

À l'inverse, dans « j'ai déjà lu ce livre », l'ordre canonique ne marque pas aussi nettement la séparation, mais le contexte discursif permet néanmoins d'inférer ce qui relève du donné et du nouveau.

En anglais, des mécanismes similaires apparaissent à travers des constructions comme « as for this book, I have already read it », où l'expression « as for » introduit explicitement le thème.

De même, l'usage de la voix passive, comme dans « the book was read by John », permet de placer en position initiale un élément déjà activé dans le discours, the book, tout en reléguant l'agent dans une position rhématique. La prédiction informationnelle repose ici sur des indices syntaxiques et discursifs qui orientent l'interprétation de la structure énonciative.

Les langues romanes telles que l'espagnol et l'italien exploitent une plus grande flexibilité de l'ordre des mots pour encoder la distinction thème rhème.

En espagnol, une phrase comme « este libro ya lo leí » place « este libro » en position thématique, renforcée par le clitique « lo » qui maintient la cohérence référentielle.

Une variation comme « ya leí este libro » peut être interprétée comme introduisant l'objet dans le rhème, selon le contexte.

En italien, « questo libro l'ho già letto » fonctionne de manière analogue, avec une dislocation à gauche et reprise pronominale. La prédiction porte ici sur la manière dont les éléments initiaux configurent l'espace des attentes pour la suite de l'énoncé.

En japonais, la distinction thème rhème est grammaticalisée de manière particulièrement explicite à travers les particules. Ainsi, la particule は, wa, marque typiquement le thème, comme dans この本はもう読んだ, kono hon wa mō yonda, ce livre, je l'ai déjà lu. L'élément marqué par « wa » est interprété comme information déjà accessible ou comme cadre de pertinence, tandis que le prédicat fournit le rhème. À l'inverse, la particule が, ga, tend à introduire des éléments nouveaux ou focalisés, comme dans 誰が来たのか, dare ga kita no ka, qui est venu, où le sujet est interrogatif et constitue le centre informatif. La prédiction énonciative en japonais repose donc sur des marqueurs morphosyntaxiques qui signalent directement la structure informationnelle.

Le chinois, en l'absence de marquage morphologique dédié, mobilise principalement l'ordre des mots et des constructions spécifiques. Une phrase comme 这本书，我已经看过了, zhè běn shū, wǒ yǐjīng kàn guò le, « ce livre, je l'ai déjà lu », utilise une structure de topicalisation où 这本书 est placé en position initiale comme thème. Le reste de l'énoncé constitue le rhème.

Dans une structure plus neutre comme 我已经看过这本书, l'objet peut être interprété comme moins présupposé, selon le contexte. La prédiction s'appuie ici sur la position initiale et sur des indices discursifs pour distinguer ce qui est donné de ce qui est nouveau.

En arabe, la structuration thème rhème peut être exprimée à travers l'ordre des constituants et des constructions spécifiques. Une phrase nominale comme الكتاب قرأته, al kitāb qara'tuhu, « le livre, je l'ai lu », place الكتاب en position initiale thématique, suivi d'un prédicat verbal avec reprise pronominale. Cette organisation permet de signaler que le référent est déjà introduit ou saillant dans le discours.

À l'inverse, une structure verbale comme قرأت الكتاب, qara'tu al kitāb, « j'ai lu le livre », peut introduire l'objet dans le rhème, selon le contexte. La prédiction énonciative doit donc intégrer des indices d'ordre, de reprise et de structure phrastique.

Au niveau cognitif, cette analyse implique que les locuteurs organisent leurs énoncés en fonction des attentes informationnelles de l'interlocuteur, en minimisant l'effort de traitement et en maximisant la pertinence. Le thème sert à ancrer l'énoncé dans un espace de référence partagé, réduisant l'incertitude initiale, tandis que le rhème apporte l'information nouvelle qui justifie l'acte de communication. La compréhension repose sur la capacité à anticiper cette organisation et à ajuster les interprétations en conséquence.

Dans le cadre des modèles de langage, la distinction thème rhème se manifeste dans la distribution des probabilités sur les séquences. Les éléments thématiques, étant plus prévisibles, présentent une probabilité conditionnelle élevée dans un contexte donné, tandis que les éléments rhématiques introduisent une variation plus importante. Les modèles apprennent ainsi à reproduire des structures informationnelles plausibles, en plaçant des éléments attendus en position initiale et en introduisant des contenus nouveaux de manière cohérente avec le contexte.

Analyse prédictive du discours

L'analyse prédictive du discours étend le cadre probabiliste à des unités supérieures à la phrase, en considérant que la production et l'interprétation des séquences discursives reposent sur des attentes structurées relatives à des schémas globaux, tels que les genres, les registres et les pratiques communicationnelles.

Le discours est ainsi conçu comme une trajectoire dans un espace de configurations possibles, où chaque segment contraint la suite en activant des modèles macroscopiques issus de l'expérience linguistique et sociale. Cette perspective permet d'articuler les dimensions locales de la cohérence avec des régularités globales propres à des types de discours stabilisés.

Au niveau interphrastique, la prédiction s'appuie sur des relations de cohérence telles que la causalité, la temporalité ou la justification.

En français, une séquence comme « il a plu toute la nuit, les routes sont inondées » active une relation causale implicite, rendant cette continuation hautement probable après la première phrase.

Une alternative comme « il a plu toute la nuit, les oiseaux chantent » reste grammaticalement acceptable, mais présente une cohérence discursive plus faible dans l'absence d'un lien contextuel explicite. La prédiction discursive repose donc sur l'activation de schémas relationnels qui structurent les enchaînements attendus entre propositions.

En anglais, des connecteurs explicites renforcent cette dynamique. Une phrase comme « she missed the train, as a result, she arrived late » guide fortement l'interprétation en signalant une relation causale.

De même, dans « a scientific article typically begins with an introduction », l'apparition du terme « introduction » active un script discursif spécifique, où l'on anticipe des sections telles que « methods, results, discussion ». La reconnaissance du genre oriente ainsi la distribution des segments attendus, au-delà du contenu lexical immédiat.

En espagnol, des connecteurs comme « entonces, sin embargo, además » signalent respectivement une conséquence, une opposition ou une addition, orientant les attentes quant à la relation entre les segments.

Dans « llegó tarde, entonces, perdió la reunión », la présence de « entonces » rend la relation causale explicite.

En italien, des expressions comme « infatti, quindi, invece » jouent un rôle analogue. La prédiction discursive intègre ici des indices lexicaux spécialisés qui encodent des relations argumentatives.

En japonais, la cohérence discursive repose en grande partie sur des particules et des formes verbales qui

indiquent les relations entre propositions. Une séquence comme 雨が降っているから、出かけない, ame ga futte iru kara, dekakenai, puisqu'il pleut, je ne sors pas, utilise « kara » pour marquer la cause, permettant d'anticiper une conséquence.

De même, la forme けど, kedo, introduit souvent une concession ou une opposition, modulant les attentes quant à la suite du discours.

La prédiction s'appuie également sur des conventions de genre, par exemple dans les récits, où l'on attend une progression temporelle structurée, souvent marquée par des formes verbales successives.

Le chinois mobilise des marqueurs discursifs et des structures parallèles pour organiser le discours. Une séquence comme 因为他生病了，所以没来, yīnwèi tā shēngbìng le, suǒyǐ méi lái, « parce qu'il était malade, il n'est pas venu », utilise 因为 et 所以 pour expliciter la relation causale.

Dans d'autres cas, l'absence de connecteurs explicites est compensée par l'ordre des propositions et par des attentes contextuelles.

Par ailleurs, certains genres discursifs, comme les récits ou les textes argumentatifs, présentent des structures récurrentes qui guident la prédiction, notamment à travers des patrons de progression thématique.

En arabe, la cohésion discursive est assurée par des connecteurs tels que لأن, li'anna, parce que, ou لكن, lākin, mais, qui orientent les relations logiques entre les segments. Une séquence comme تأخر لأنه مريض, ta'akhkhara li'annahu marīḍ, « il a été en retard parce qu'il est malade », illustre une relation causale explicitée.

Dans des textes plus longs, la répétition lexicale et l'usage de structures parallèles contribuent également à la cohérence, permettant au lecteur d'anticiper les développements à venir.

Au-delà des relations locales, l'analyse prédictive du discours met en évidence le rôle des genres comme cadres contraignants. Un article scientifique, un récit narratif, une conversation informelle ou un discours politique mobilisent chacun des structures attendues, qui orientent la production et l'interprétation.

Par exemple, dans un article scientifique en français ou en anglais, on anticipe une organisation en introduction, méthodologie, résultats et discussion, tandis qu'un récit en japonais ou en chinois suit souvent une progression narrative structurée autour d'événements successifs. Ces schémas sont appris à partir de l'exposition à des corpus et constituent des modèles internes guidant la prédiction.

Dans le cadre des grands modèles de langage, cette dimension se traduit par l'apprentissage de régularités à grande échelle sur des corpus variés. Les modèles internalisent des structures discursives associées à différents genres, ce qui leur permet de produire des textes cohérents sur plusieurs phrases, en respectant des attentes globales.

Cette capacité repose sur la modélisation conjointe de dépendances locales et de contraintes globales, captant ainsi des schémas récurrents dans les données.

Méthodologie de l'analyse prédictive

La méthodologie de l'analyse prédictive combine l'exploitation de corpus à grande échelle, l'annotation de phénomènes discursifs et l'utilisation de modèles computationnels capables d'estimer des distributions conditionnelles sur des séquences étendues. L'objectif n'est pas seulement de décrire des structures existantes, mais d'identifier les contraintes qui rendent certaines continuations discursives plus probables que d'autres dans un contexte donné.

Une première étape consiste en la constitution et la structuration de corpus représentatifs de différents genres et registres.

En français, un corpus peut inclure des articles scientifiques, des conversations orales et des textes journalistiques, permettant d'observer des variations dans l'usage des connecteurs comme « donc, cependant, en effet… ».

En anglais, des corpus comparables permettent d'analyser la distribution de marqueurs tels que « therefore, however, indeed… », et leur rôle dans l'organisation argumentative.

La diversité des données est essentielle pour modéliser la variabilité des attentes discursives selon les contextes.

L'annotation constitue une étape centrale, visant à expliciter les relations discursives, les structures informationnelles et les indices pragmatiques.

En espagnol, une séquence comme « llegó tarde, entonces perdió la reunión » peut être annotée comme une relation de conséquence, de même qu'en italien, « è arrivato tardi, quindi ha perso la riunione » présente une configuration analogue.

En japonais, 雨が降っているから、出かけない peut être annoté comme une relation causale marquée par から, kara.

En chinois, 因为他生病了，所以没来 offre un cas où la relation est doublement marquée, facilitant

l'annotation.

En arabe, تأخر لأنه مريض permet d'identifier explicitement le lien causal introduit par لأن. Ces annotations servent de base à l'apprentissage et à l'évaluation des modèles prédictifs.

Une fois les corpus annotés, des modèles statistiques ou neuronaux sont entraînés pour estimer la probabilité de segments discursifs conditionnellement à leur contexte. Par exemple, en français, après une phrase comme « il a raté son train », un modèle attribuera une probabilité élevée à une continuation telle que « il est arrivé en retard », en raison de la relation causale fréquemment observée dans les données.

En anglais, après « she studied all night », une suite comme « she passed the exam » sera jugée plus probable que « she forgot her name », sauf contexte particulier. Ces modèles capturent des régularités qui dépassent les relations syntaxiques locales pour inclure des schémas de cohérence à plus grande échelle.

La méthodologie inclut également l'analyse des genres discursifs comme cadres prédictifs. Dans un article scientifique en français ou en anglais, la présence d'une introduction augmente la probabilité d'une section méthodologique, puis de résultats.

En espagnol ou en italien, des structures similaires apparaissent dans les textes académiques.

En japonais, un récit narratif suit souvent une progression temporelle marquée par des formes verbales successives, tandis qu'en chinois, les textes argumentatifs peuvent présenter des structures parallèles et des répétitions lexicales.

En arabe, les discours formels peuvent inclure des séquences rhétoriques attendues, comme des formules d'ouverture et de clôture. L'identification de ces schémas permet de modéliser des attentes à un niveau macrostructural.

Une dimension importante de cette méthodologie concerne l'évaluation des prédictions. Celle-ci peut être réalisée en comparant les sorties des modèles à des données réelles, ou en mesurant la capacité à anticiper correctement des segments manquants.

Par exemple, dans une tâche de complétion de texte en français, on peut supprimer un connecteur comme « cependant » et évaluer si le modèle propose une continuation cohérente avec une relation d'opposition.

Des expériences analogues peuvent être menées en anglais, espagnol, italien, japonais, chinois et arabe, afin de tester la robustesse des modèles face à des structures discursives variées.

Enfin, la méthodologie de l'analyse prédictive du discours implique une réflexion sur l'interprétabilité des modèles et sur leur capacité à rendre compte des phénomènes linguistiques observés.

Si les modèles peuvent apprendre des régularités complexes à partir de corpus, il reste nécessaire de relier ces régularités à des catégories descriptives, telles que les relations discursives, les structures informationnelles ou les intentions pragmatiques. Cette articulation entre données, modèles et théorie constitue un enjeu central pour le développement d'une approche intégrée du discours.

Applications de l'analyse prédictive

Les applications de l'analyse prédictive du discours se déploient dans un ensemble de domaines où la maîtrise des structures discursives et des attentes contextuelles constitue un levier pour améliorer la compréhension, la génération et l'interprétation du langage. En s'appuyant sur des régularités probabilistes apprises à partir de corpus, ces applications exploitent la capacité à anticiper non seulement des unités linguistiques locales, mais aussi des trajectoires discursives complètes, adaptées à des genres, des contextes et des intentions spécifiques.

Dans le domaine de la génération automatique de texte, l'analyse prédictive du discours permet de produire des séquences cohérentes sur plusieurs phrases, en respectant les contraintes propres à différents genres.

En français, un système peut générer un article d'actualité en structurant l'information selon un schéma attendu, avec une introduction factuelle suivie de développements explicatifs.

En anglais, la production d'un résumé scientifique peut suivre une organisation canonique où une phrase initiale contextualise l'étude, suivie d'une présentation des résultats principaux.

En espagnol et en italien, des systèmes analogues peuvent générer des textes argumentatifs en mobilisant des connecteurs comme « entonces » ou « quindi » pour structurer les relations logiques.

Dans le champ de la traduction automatique, la prise en compte des structures discursives améliore la fidélité et la fluidité des traductions. Par exemple, une relation causale explicitée en chinois par 因为 et 所以 peut être rendue en français par « parce que » et « donc », ou parfois par une structure implicite si le contexte le permet.

En japonais, une phrase comportant けど, kedo, peut nécessiter une traduction en anglais ou en français qui explicite ou atténue l'opposition selon les normes

discursives de la langue cible.

En arabe, des connecteurs comme لكن, lākin, doivent être interprétés non seulement comme des marqueurs d'opposition, mais aussi comme des indices de progression argumentative. L'analyse prédictive du discours permet d'aligner ces structures au-delà du niveau phrastique.

Les systèmes de résumé automatique bénéficient également de cette approche, en identifiant les segments discursifs les plus informatifs et en préservant la cohérence globale.

En français, dans un texte narratif, les événements clés peuvent être sélectionnés en fonction de leur rôle dans la progression temporelle.

En anglais, dans un article d'opinion, les arguments principaux et les conclusions sont privilégiés.

En chinois, la détection de structures parallèles et de répétitions lexicales peut guider la condensation du texte, tandis qu'en japonais, la hiérarchie des informations peut être inférée à partir de la structure des propositions et des marqueurs discursifs.

Dans l'analyse et la classification des genres, l'approche prédictive permet d'identifier automatiquement le type de discours en fonction de ses caractéristiques globales. Un texte en français contenant des marqueurs comme « en conclusion » ou « en effet », associé à une structure argumentative, sera classé différemment d'un récit narratif.

En anglais, la présence de sections telles que « introduction », « methods » ou « discussion » signale un article scientifique.

En espagnol et en italien, des indices lexicaux et structurels analogues permettent de distinguer des genres académiques, journalistiques ou conversationnels.

En arabe, des formules d'ouverture et de clôture peuvent indiquer un discours formel ou religieux.

Les applications en interaction homme-machine exploitent la capacité à anticiper les intentions discursives pour améliorer la pertinence des réponses. Dans un dialogue en français, une question indirecte comme « tu sais quelle heure il est » peut être interprétée comme une demande d'information temporelle, même si elle n'est pas formulée de manière directe.

En anglais, une phrase comme « I was wondering if you could help me » déclenche une attente de requête.

En japonais, une formulation atténuée peut nécessiter une interprétation pragmatique fine pour produire une réponse appropriée.

En chinois, des réponses brèves comme 可以 doivent être contextualisées pour déterminer leur portée. L'analyse prédictive du discours permet d'intégrer ces indices pour ajuster les interactions.

Dans le domaine de l'enseignement des langues, cette approche offre des outils pour modéliser et enseigner les structures discursives propres à chaque langue.

En français, l'apprentissage des connecteurs logiques et de la structuration argumentative peut être guidé par des modèles prédictifs qui illustrent les enchaînements typiques.

En anglais, la maîtrise des genres académiques repose sur la compréhension de schémas discursifs récurrents.

En espagnol et en italien, l'usage des marqueurs discursifs peut être enseigné à partir de leur distribution dans des corpus.

En japonais et en chinois, l'accent peut être mis sur les structures informationnelles et les conventions pragmatiques, tandis qu'en arabe, l'apprentissage inclut des formes rhétoriques spécifiques.

Enfin, dans l'analyse des corpus et la recherche linguistique, l'approche prédictive permet de mettre au jour des régularités discursives à grande échelle, en comparant des langues et des genres. Elle offre des outils pour explorer la variation, détecter des anomalies ou modéliser des évolutions diachroniques. En combinant des données multilingues, elle contribue à une compréhension plus fine des mécanismes de cohérence et de structuration du discours.

Application à la santé mentale

Les productions verbales, orales ou écrites, contiennent des indices structuraux et sémantiques permettant d'anticiper des états psychiques et leurs évolutions. Dans cette perspective, le discours est envisagé comme une trace dynamique de l'activité cognitive et émotionnelle, dont les régularités et les écarts peuvent être modélisés afin d'identifier des profils à risque ou des changements cliniquement pertinents.

Sur le plan méthodologique, l'approche consiste à entraîner des modèles sur des corpus de récits associés à des états cliniques connus, afin d'estimer la probabilité de certaines configurations discursives.

En français, des récits marqués par une forte récurrence de formes négatives, comme « rien ne va, je n'y arrive pas », ou par une réduction de la diversité lexicale, peuvent être associés à des états dépressifs.

La prédiction ne repose pas sur un indice isolé, mais sur la combinaison de traits tels que la structure

narrative, la cohérence globale et la distribution des thèmes abordés.

En anglais, des expressions comme « I feel empty » ou « nothing matters » peuvent apparaître dans des contextes similaires, mais leur interprétation dépend de leur fréquence relative et de leur insertion dans le discours.

En espagnol et en italien, des phénomènes analogues peuvent être observés, avec des variations liées aux conventions discursives.

En espagnol, un récit comportant des séquences comme « no tengo ganas de nada » ou « todo es inútil » peut signaler une tonalité affective négative persistante.

En italien, des expressions comme « non ha senso » ou « sono stanco di tutto » peuvent jouer un rôle comparable.

L'analyse prédictive intègre également des indices temporels, tels que l'usage dominant du passé ou du présent, qui peuvent refléter des orientations cognitives différentes, par exemple une rumination centrée sur des événements passés.

Dans des langues comme le japonais, où l'implicite et l'atténuation jouent un rôle important, les indices peuvent être plus subtils. Une expression comme ちょっと疲れました, chotto tsukaremashita, « je suis un peu fatigué », peut, selon le contexte et la fréquence, signaler un état de détresse plus profond que ne le suggère la traduction littérale. L'omission du sujet et la structure fragmentaire de certains énoncés peuvent également constituer des indices d'un retrait discursif.

En chinois, des formulations comme 没什么意思, méi shénme yìsi, cela n'a pas beaucoup de sens, ou 我不想说话, wǒ bù xiǎng shuōhuà, je ne veux pas parler, peuvent être interprétées à la lumière de leur distribution et de leur évolution dans le temps.

L'arabe présente également des marqueurs discursifs pertinents, où des expressions telles que لا فائدة, lā fā'ida, « cela ne sert à rien », ou des répétitions de structures exprimant la résignation peuvent signaler des états de détresse.

La dimension culturelle est ici déterminante, car certaines formes d'expression de la souffrance peuvent être indirectes ou ritualisées, ce qui nécessite une modélisation adaptée aux normes discursives locales.

Au-delà des indices lexicaux, l'analyse prédictive du discours en santé mentale prend en compte des caractéristiques structurelles plus globales. La désorganisation narrative, par exemple, peut se manifester par des ruptures de cohérence, des transitions abruptes ou une difficulté à maintenir un fil thématique.

À l'inverse, une rigidité excessive dans la répétition de certains schémas peut également constituer un signal. Dans des récits en français ou en anglais, une focalisation persistante sur des thèmes négatifs, sans variation ou élaboration, peut être modélisée comme une distribution discursive atypique.

Les modèles prédictifs permettent également de suivre l'évolution temporelle des discours. En analysant des séries de textes produits par une même personne, il est possible d'identifier des changements progressifs, comme une augmentation de la négativité, une simplification syntaxique ou une modification des relations discursives. Ces variations peuvent être observées dans différentes langues, bien que leur manifestation dépende des संसोर्स linguistiques disponibles. Par exemple, en espagnol ou en italien, des changements dans l'usage des temps verbaux ou des connecteurs peuvent refléter des évolutions dans la structuration cognitive du récit.

Dans les applications cliniques, ces approches peuvent être intégrées à des outils de dépistage ou de suivi, en complément des évaluations traditionnelles. Un système analysant des messages écrits en français, en anglais ou en chinois pourrait signaler des configurations discursives associées à un risque accru, permettant une intervention précoce. Dans des contextes multilingues, la prise en compte des spécificités linguistiques est essentielle pour éviter des biais d'interprétation.

Application à l'analyse des émotions

Les états affectifs se manifestent à travers des configurations linguistiques récurrentes, dont la distribution peut être modélisée et anticipée. En effet, les émotions ne sont pas uniquement exprimées par des marqueurs lexicaux explicites, mais émergent de l'interaction entre choix lexicaux, structures syntaxiques, organisation discursive et indices pragmatiques. L'approche prédictive vise ainsi à estimer la probabilité de certaines tonalités émotionnelles à partir de patrons observés dans les données, en intégrant la variabilité contextuelle.

Au niveau lexical, certains champs sémantiques sont associés à des états affectifs particuliers, mais leur interprétation dépend du contexte.

En français, des mots comme « heureux, inquiet, épuisé » peuvent signaler respectivement des états positifs, anxieux ou dépressifs, mais leur portée émotionnelle est modulée par leur insertion discursive. Une phrase comme « je suis fatigué » peut

renvoyer à un état physique neutre ou à une lassitude plus profonde selon le cotexte.

En anglais, des expressions comme « I'm fine » peuvent, dans certains contextes, dissimuler une émotion négative, ce qui nécessite une analyse au-delà du sens littéral.

La prédiction émotionnelle implique donc une pondération des indices explicites et implicites.

Dans les langues romanes comme l'espagnol et l'italien, la morphologie et les marqueurs discursifs contribuent à la modulation affective.

En espagnol, « estoy muy cansado » peut exprimer une fatigue intense, mais l'ajout de « siempre », comme dans « siempre estoy cansado », introduit une dimension de persistance qui peut être associée à un état émotionnel plus problématique.

En italien, « sono contento » et « sono proprio contento » diffèrent par le degré d'intensité, la présence de « proprio » renforçant l'expression émotionnelle. La prédiction doit intégrer ces modulateurs qui affectent la polarité et l'intensité.

Dans des langues comme le japonais, l'expression des émotions est souvent indirecte et contextualisée. Une phrase comme なんだか寂しい, nandaka sabishii, je me sens un peu seul, utilise un adverbe atténuateur qui nuance l'intensité de l'émotion.

De même, l'usage de particules finales comme ね, ne, peut signaler une recherche de partage émotionnel avec l'interlocuteur. L'absence de sujet explicite et la dépendance au contexte rendent l'inférence plus complexe, nécessitant une modélisation fine des attentes interactionnelles.

En chinois, des expressions comme 有点难过, yǒudiǎn nánguò, un peu triste, ou 很开心, hěn kāixīn, très heureux, illustrent des degrés d'intensité marqués par des adverbes. Toutefois, des énoncés comme 还行, hái xíng, ça va, peuvent masquer des émotions négatives selon le contexte, ce qui requiert une analyse discursive globale.

L'arabe offre également des indices spécifiques, où l'intensité émotionnelle peut être exprimée par la répétition ou par des constructions emphatiques. Une expression comme أنا حزين جدا, anā ḥazīn jiddan, « je suis très triste », combine un adjectif et un intensificateur, tandis que des formulations plus indirectes peuvent signaler des émotions à travers des métaphores ou des expressions idiomatiques. La prédiction émotionnelle doit ici tenir compte des normes culturelles d'expression et des variations dialectales.

Au niveau discursif, les émotions se manifestent par des configurations plus larges, incluant la structure narrative, la cohérence et la progression thématique. Ainsi, un récit en français marqué par des ruptures fréquentes, des répétitions ou une focalisation sur des événements négatifs peut être associé à une tonalité émotionnelle particulière.

En anglais, une alternance entre des évaluations positives et négatives peut signaler une ambivalence affective.

En espagnol ou en italien, l'usage récurrent de connecteurs comme « pero » ou « ma » peut indiquer des contrastes émotionnels internes. La prédiction s'appuie ici sur des patrons discursifs qui dépassent les unités isolées.

Les modèles computationnels exploitent ces régularités en associant des représentations vectorielles aux segments de discours, permettant de capter des nuances émotionnelles à partir de leur distribution contextuelle. En analysant de larges corpus multilingues, ils apprennent à corréler certaines configurations linguistiques avec des catégories émotionnelles ou des dimensions continues telles que la valence et l'intensité. Cette capacité permet des applications telles que la détection d'émotions dans des interactions en ligne ou l'analyse de retours utilisateurs dans différentes langues.

Application à l'analyse de la culture

Les pratiques culturelles, les valeurs sociales et les représentations collectives se manifestent à travers des régularités linguistiques observables dans les corpus textuels. Les mots, les expressions et les structures discursives ne sont pas distribués aléatoirement, mais apparaissent selon des configurations qui reflètent des cadres culturels partagés. L'approche prédictive vise ainsi à modéliser ces associations en estimant la probabilité qu'un phénomène culturel soit évoqué, décrit ou évalué à travers certaines formes linguistiques.

Au niveau lexical, certaines unités sont fortement indexées sur des pratiques culturelles spécifiques. En français, le terme « laïcité » apparaît fréquemment dans des contextes liés à l'école, à l'État ou au débat public, ce qui permet d'anticiper des thématiques institutionnelles et politiques.

En anglais, des expressions comme « freedom of speech » ou « individual responsibility » sont souvent associées à des discours valorisant l'autonomie individuelle.

En espagnol, « fiesta » peut évoquer non seulement une célébration, mais aussi des pratiques sociales

spécifiques liées à des contextes locaux.

En italien, « famiglia » est fréquemment mobilisé dans des contextes valorisant les relations interpersonnelles et les structures familiales élargies. La prédiction culturelle repose ici sur des réseaux de cooccurrences qui ancrent les mots dans des univers de sens collectifs.

Dans des langues comme le japonais, certaines expressions encapsulent des normes sociales spécifiques. Le terme 空気を読む, kūki o yomu, littéralement « lire l'air », est utilisé pour décrire la capacité à percevoir implicitement les attentes sociales, ce qui reflète une valorisation de l'harmonie et de l'ajustement contextuel. La présence de cette expression dans un discours permet d'anticiper des thématiques liées à la conformité sociale ou à la gestion des interactions.

En chinois, des concepts comme 面子, miànzi, « la face », apparaissent dans des contextes où la réputation et l'honneur social sont en jeu. Leur occurrence oriente l'interprétation vers des enjeux relationnels et symboliques spécifiques.

L'arabe offre également des exemples où certaines expressions sont étroitement liées à des cadres culturels. Des formules comme السلام عليكم, as salāmu ʿalaykum, ou الحمد لله, al ḥamdu li llāh, ou encore « inchallah », apparaissent dans des contextes interactionnels et religieux, signalant des normes de politesse et des références spirituelles. Leur fréquence et leur distribution permettent de modéliser des pratiques discursives ancrées dans des traditions culturelles. La prédiction ne porte pas seulement sur le contenu lexical, mais sur les situations dans lesquelles ces expressions sont attendues.

Au niveau discursif, les genres et les structures narratives reflètent également des modèles culturels. En français ou en anglais, un discours politique peut mobiliser des oppositions binaires et des arguments structurés, tandis qu'en chinois, des textes argumentatifs peuvent recourir à des parallélismes et à des structures répétitives.

En japonais, les récits peuvent privilégier une progression implicite et une attention aux relations interpersonnelles.

En arabe, certains discours formels incluent des séquences rhétoriques spécifiques, telles que des formules d'ouverture et des invocations. L'analyse prédictive permet d'identifier ces schémas et d'anticiper les structures discursives en fonction du contexte culturel.

Les modèles computationnels exploitent ces régularités en apprenant des associations entre formes linguistiques et contextes culturels à partir de grands corpus. Par exemple, en analysant des textes en plusieurs langues, un modèle peut identifier que des termes liés à la nourriture, comme « fromage » en français, « pasta » en italien ou 饺子, jiǎozi, en chinois, apparaissent dans des contextes spécifiques liés à des pratiques culinaires. Ces associations permettent de générer ou d'interpréter des discours en tenant compte de cadres culturels implicites.

Cette approche trouve des applications dans la traduction et l'adaptation culturelle, où il est nécessaire de préserver non seulement le sens littéral, mais aussi les connotations culturelles. Une expression idiomatique en anglais peut nécessiter une reformulation en français ou en espagnol pour maintenir une équivalence pragmatique. De même, des références culturelles explicites dans un texte chinois ou japonais peuvent être adaptées pour un public différent. L'analyse prédictive du discours fournit des outils pour identifier ces correspondances et ajuster les productions en conséquence.

Cependant, l'inférence culturelle à partir du langage comporte des limites. En effet, les associations observées dans les corpus peuvent refléter des biais ou des stéréotypes, et leur généralisation doit être abordée avec prudence. La diversité interne des cultures et la variation contextuelle impliquent que les régularités statistiques ne captent qu'une partie de la complexité des pratiques culturelles. Il est donc nécessaire de compléter les analyses quantitatives par des approches qualitatives et contextualisées.

Lexique et définitions

La linguistique prédictive constitue un cadre conceptuel unifié, permettant de décrire le langage comme un système d'anticipation à plusieurs niveaux.

Probabilité conditionnelle

Désigne la probabilité d'apparition d'une unité linguistique donnée en fonction du contexte qui la précède. Ce principe constitue le fondement formel de la prédiction dans les modèles de langage.

Distribution linguistique

Ensemble des occurrences d'une unité linguistique dans différents contextes. Elle permet d'inférer ses propriétés syntaxiques, sémantiques et pragmatiques.

Modèle de langage

Système computationnel qui estime la probabilité de séquences linguistiques, généralement en prédisant un élément à partir de son contexte.

Token

Unité de traitement dans un modèle, pouvant correspondre à un mot, une sous partie de mot ou un symbole. Les tokens constituent les éléments de base de la prédiction.

Contexte

Ensemble des unités précédentes et des informations associées qui conditionnent l'interprétation et la production d'une unité linguistique.

Fenêtre contextuelle

Portion de texte effectivement prise en compte par un modèle pour effectuer une prédiction.

Perplexité

Mesure de la qualité d'un modèle de langage, indiquant dans quelle mesure celui-ci est surpris par les phénomènes observés. Une perplexité faible indique une meilleure capacité prédictive.

Entropie linguistique

Mesure du degré d'incertitude associé à une distribution de probabilités sur des unités linguistiques.

Surprisalité

Quantification de l'imprévisibilité d'une unité donnée dans un contexte. Plus une unité est inattendue, plus sa surprisalité est élevée.

Apprentissage non supervisé

Mode d'apprentissage dans lequel le modèle extrait des régularités à partir de données non annotées, en optimisant un objectif prédictif.

Apprentissage auto supervisé

Forme d'apprentissage où les algorithmes eux-mêmes fournissent les cibles d'entraînement, par exemple en masquant ou en prédisant des éléments.

Représentation distribuée

Encodage vectoriel des unités linguistiques, où le sens émerge de la position relative dans un espace multidimensionnel.

Embedding

Vecteur numérique représentant une unité linguistique, en captant ses traits sémantiques et ses relations avec d'autres unités.

Similarité distributionnelle

Principe selon lequel des unités apparaissant dans des contextes similaires ont des propriétés similaires.

Modélisation séquentielle

Approche consistant à traiter le langage comme une séquence ordonnée d'unités dépendantes les unes des autres.

Dépendances à longue distance

Relations entre unités linguistiques séparées par plusieurs positions dans la séquence.

Transformeur

Architecture neuronale fondée sur des mécanismes d'attention, permettant de modéliser efficacement les relations globales.

Attention

Mécanisme permettant de pondérer l'importance relative des éléments du contexte lors de la prédiction.

Attention multi têtes

Extension du mécanisme d'attention permettant de capter différents types de relations simultanément.

Encodage positionnel

Procédé permettant d'introduire l'information d'ordre dans les modèles qui traitent les unités de manière parallèle.

Génération autorégressive

Processus de production séquentielle où chaque unité est générée à partir des précédentes.

Décodage

Étape de génération des sorties linguistiques à partir des représentations internes du modèle.

Température

Paramètre contrôlant le degré de variation dans la génération, influençant la diversité des sorties.

Grammaticalité

Conception selon laquelle la grammaticalité d'un énoncé est graduelle et dépend de sa probabilité dans un contexte donné.

Acceptabilité

Jugement linguistique relatif à la naturalité d'un énoncé, modélisé comme une variable continue.

Prédiction syntaxique

Anticipation des structures grammaticales à partir des régularités observées dans les phrases du discours.

Prédiction sémantique

Inférence du sens d'une unité ou d'un énoncé en fonction du contexte.

Prédiction pragmatique

Anticipation des intentions communicatives et des effets discursifs.

Prédiction morphologique

Analyse des unités sublexicales pour anticiper la structure et le sens des mots.

Prédiction discursive

Anticipation des enchaînements au-delà de la phrase, incluant la cohérence et les relations entre segments.

Structure informationnelle

Organisation des énoncés en termes de thème et de rhème, liée à la gestion de l'information.

Thème

Partie de l'énoncé correspondant à l'information déjà connue ou accessible.

Rhème

Partie apportant une information nouvelle ou moins prévisible.

Cohérence

Propriété d'un texte dont les différentes parties sont reliées par des relations interprétables.

Cohésion

Ensemble des moyens linguistiques assurant les liens entre les éléments du discours.

Genre

Type de discours caractérisé par des structures et des attentes spécifiques.

Script discursif

Schéma cognitif représentant une structure typique d'événements ou d'énoncés.

Inférence pragmatique

Processus par lequel l'interlocuteur déduit des informations non explicitement exprimées.

Contexte partagé

Ensemble des connaissances supposées communes aux participants d'une interaction.

Ancrage contextuel

Processus d'intégration d'un énoncé dans une situation donnée.

Ambiguïté

Présence de plusieurs interprétations possibles pour une même forme linguistique.

Désambiguïsation

Processus de sélection de l'interprétation la plus probable en contexte.

Polysémie

Propriété d'un mot ayant plusieurs sens interreliés.

Métaphore

Processus de transfert de sens basé sur une analogie entre domaines.

Représentation latente

Structure interne du modèle qui encode des informations non directement observables.

Émergence

Apparition de propriétés complexes à partir de règles simples d'apprentissage.

Généralisation

Capacité d'un modèle à appliquer des régularités apprises à de nouveaux cas.

Biais de corpus

Distorsion introduite par la distribution des données utilisées pour l'entraînement.

Alignement culturel

Processus visant à adapter les sorties d'un modèle aux attentes humaines.

Interprétabilité

Capacité à comprendre le fonctionnement interne d'un modèle de langage.

Évaluation intrinsèque

Mesure des performances basée sur des critères internes comme la perplexité.

Évaluation extrinsèque

Mesure des performances dans des tâches appliquées.

Robustesse

Capacité d'un modèle à maintenir ses performances face à des corpus variés ou bruités.

Transfert

Utilisation de connaissances acquises dans un contexte pour une autre tâche ou langue.

Multilinguisme

Capacité d'un modèle à traiter plusieurs langues et à exploiter des régularités communes.

Annexe : Liste des modèles de langage analysés

Chaque modèle est envisagé comme un dispositif d'apprentissage et de modélisation des régularités linguistiques, discursives et culturelles propres à des écosystèmes spécifiques.

La Chine

Dans le contexte chinois, les modèles de langage s'inscrivent dans une dynamique de modélisation prédictive fortement contrainte par des normes institutionnelles. Ils apprennent à anticiper non seulement des structures syntaxiques et sémantiques, mais aussi des configurations discursives compatibles avec des cadres sociopolitiques spécifiques.

- *Baidu avec Ernie Bot*

Ce modèle repose sur une optimisation des probabilités conditionnelles dans des corpus sinophones massifs, intégrant des régularités lexicales, syntaxiques et pragmatiques propres au mandarin standard. Il est particulièrement orienté vers la prédiction de structures discursives informatives et institutionnelles.

• *Alibaba avec Qwen*

Qwen se distingue par une modélisation prédictive adaptée aux discours commerciaux et transactionnels. Il apprend à anticiper des séquences linguistiques typiques des interactions économiques, en exploitant des distributions lexicales et discursives issues du commerce numérique.

• *Zhipu AI*

Issu d'un environnement académique, ce modèle met l'accent sur la formalisation des relations linguistiques complexes, notamment dans des contextes scientifiques et techniques, où la précision des prédictions syntaxiques et sémantiques est centrale.

• *DeepSeek*

Ce modèle illustre une optimisation efficace de la prédiction linguistique à grande échelle, en maximisant la performance tout en minimisant le coût computationnel. Il met en évidence la possibilité de capter des régularités discursives complexes avec des moyens limités.

La Russie

Les modèles russes s'inscrivent dans une logique de souveraineté linguistique, en apprenant des distributions spécifiques à la langue russe et à ses usages discursifs, notamment dans les domaines institutionnels et médiatiques.

• *Yandex avec YandexGPT*

Ce modèle est intégré à un écosystème numérique local, ce qui lui permet de modéliser finement les pratiques discursives russophones, en anticipant des structures propres aux interactions quotidiennes et aux services numériques.

• *Sberbank avec GigaChat*

GigaChat est orienté vers des usages institutionnels et économiques, avec une capacité à prédire des discours liés aux services financiers, en intégrant des contraintes terminologiques et pragmatiques spécifiques.

Le Moyen Orient

Dans cette région, les modèles de langage sont conçus pour capter les spécificités de l'arabe, tant dans sa variation dialectale que dans ses usages formels, ainsi que les cadres culturels associés.

• *Falcon*

Falcon illustre une modélisation prédictive de haut niveau, capable de capter des dépendances linguistiques complexes en arabe et en anglais. Son ouverture favorise l'analyse comparative des régularités discursives entre langues.

• *Jais*

Jais est explicitement orienté vers la modélisation des distributions linguistiques reflétant la culture du Golfe, en intégrant des patrons discursifs et pragmatiques propres aux contextes arabophones.

• *Fanar*

Ce modèle privilégie la prédiction en arabe standard, en se concentrant sur des corpus spécialisés, ce qui permet une modélisation fine des discours techniques et institutionnels.

• *Humain*

Ce modèle multimodal étend la prédiction au-delà du texte, en intégrant des données visuelles. Il apprend des corrélations entre langage, image et contexte sectoriel, notamment dans des domaines de l'énergie ou de la santé.

L'Inde

Les modèles indiens sont caractérisés par une approche multilingue, visant à modéliser des distributions linguistiques très hétérogènes.

• *Krutrim*

Ce modèle est conçu pour apprendre des régularités à travers plusieurs langues indiennes, en captant des correspondances interlinguistiques et des variations morphosyntaxiques.

• *Bhashini*

Bhashini vise à modéliser la diversité linguistique comme un espace prédictif unifié, permettant de transférer des connaissances entre langues et de réduire les barrières communicationnelles.

La Corée du Sud

Les modèles coréens mettent l'accent sur la précision des prédictions dans des contextes socioculturels spécifiques.

• *Naver avec HyperCLOVA X*

Ce modèle capte des phénomènes linguistiques liés aux normes sociales, juridiques et interactionnelles du coréen, en intégrant des contraintes pragmatiques fines.

L'Europe

Les modèles européens se distinguent par une attention particulière à la diversité linguistique et aux cadres réglementaires.

• *Mistral AI*

Ce modèle illustre une approche visant à modéliser des distributions linguistiques multilingues, tout en respectant des critères éthiques et juridiques spécifiques au contexte européen.

Références

Barbazan, M. (2008). Principes d'une grammaire prédictive du discours (français langue étrangère et maternelle). In *Congrès Mondial de Linguistique Française* (p. 053). EDP Sciences.

Barbazan, M. (2010). Modèles explicatifs, modèles prédictifs: pour une interaction effective entre linguistique et cognition. In *Temps, aspect et modalité en français* (pp. 25-43). Brill Rodopi.

Bazziconi, P. F., Berrouiguet, S., Kim-Dufor, D. H., Walter, M., & Lemey, C. (2021). Les marqueurs linguistiques dans l'amélioration du modèle prédictif de la transition vers la schizophrénie. *L'Encéphale*, *47*(5), 499-501.

Bazziconi, P. F., Bleton, L., Berrouiguet, S., Thierry, A., Walter, M., & Lemey, C. (2019). L'utilisation de marqueurs linguistiques et de méthodes d'apprentissage automatique du discours dans la prédiction de la transition vers la psychose: quels enjeux pour le patient et le psychiatre?. *L'information psychiatrique*, *95*(2), 89-94.

Bernelin, M. (2022). L'intelligence artificielle dans le domaine de la justice: réflexions sur la «représentation sémantique» du droit. *Penser, calculer, délibérer (Ph. PEDROT & Alain PAPAUX dir.)*.

Bjerva, J. (2023). The role of typological feature prediction in NLP and linguistics. *Computational Linguistics*, *50*(2), 781-794.

Blache, P. (2025). La prédiction au cœur de la communication : une perspective interdisciplinaire entre linguistique, intelligence artificielle et neurosciences. *Histoire (s) en mouvement*, 177-193.

Bottemanne, H., Longuet, Y., & Gauld, C. (2022). L'esprit predictif: introduction à la théorie du cerveau bayésien. *L'Encéphale*, *48*(4), 436-444.

Choi, H. S., Trivedi, P., Constant, M., Fort, K., & Guillaume, B. (2024). Au-delà de la performance des modèles: la prédiction de liens peut-elle enrichir des graphes lexico-sémantiques du français?. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1: articles longs et prises de position* (pp. 36-49).

Churunina, A. A., Solnyshkina, M. I., & Yarmakeev, I. E. (2023). Lexical diversity as a predictor of complexity in textbooks on the Russian language. *Russian Language Studies*, *21*(2), 212-227.

Gavard, E. (2024). *Le rôle de la prédiction sémantique et syntaxique et de l'apprentissage statistique dans la lecture normale et dans la dyslexie développementale* (Doctoral dissertation, Aix-Marseille Université).

Grabar, N., & Eshkol, I. (2016). Prédiction automatique de fonctions pragmatiques dans les reformulations (Automatic prediction of pragmatic functions in reformulations). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2: TALN (Articles longs)* (pp. 262-275).

Guidère, M. (2015). La linguistique prédictive: de la cognition à l'action. Paris : Éditions L'Harmattan.

Guidère, M. (2024). Predictive Linguistics, Inner Language, Mental Health. *Journal of Applied Research in Human and Social Sciences*, *1*(1), 27-42.

Huang, Z., Hu, X., & Jin, H. (2025). The Predictive Effects of L2 Writing Anxiety on Motivational Regulation Strategies: A Person-Centered Approach. *International Journal of Applied Linguistics*, *35*(3), 1045-1057.

Hubert, N. (2024). *From semantic-aware to semantic-enhanced knowledge graph embedding models for link prediction* (Doctoral dissertation, Université de Lorraine).

Liu, T., Jiang, Y. E., Monath, N., Cotterell, R., & Sachan, M. (2022, December). Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 993-1005).

Liu, Y., Zhang, X. F., Wegsman, D., Beauchamp, N., & Wang, L. (2022, July). POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 1354-1374).

Mertens, P. (2008). Syntaxe, prosodie et structure informationnelle : une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de linguistique*, *56*(1), 97-124.

Peper, J. J., & Wang, L. (2022, December). Generative aspect-based sentiment analysis with contrastive learning and expressive structure. In *Findings of the association for computational linguistics: EMNLP 2022* (pp. 6089-6095).

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: what is next?. *Trends in cognitive sciences*, *27*(11), 1032-1052.

Troiano, E., Oberländer, L., & Klinger, R. (2023). Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, *49*(1), 1-72.

Zou, M., Teng, M. F., Soyoof, A., & He, X. (2026). Informal Digital Learning of English (IDLE) as form-focused and meaning-focused activities: Refining its measurement and examining its predictive role in L2 achievement and confidence. *International Journal of Applied Linguistics*, *36*(1), 393-408.

Brevet associé

Patent: Method for Cognitive Computing

Brevet: Méthode de programmation cognitive

Inventor:(Mathieu Guidere) :

https://patentscope.wipo.int/search/fr/detail.jsf?docId=US73511719&_cid=P22-K269BG-27302-1