Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

De la conception du formalisme de représentation sémantique MR4AP aux solutions métiers

Cédric Lopez Lucas Aubertin Jean Bort Stéphane Froliger
(1) Emvista R&D, 10 rue Louis Breguet, 34830 Jacou, France
prenom.nom@emvista.com

RÉSUMÉ

Cet article retrace les étapes principales du développement de la technologie de structuration de l'information textuelle Prevyo, depuis la conception du formalisme MR4AP jusqu'à sa mise en œuvre dans un système opérationnel. À travers cet article, nous partageons l'intuition de l'utilité de MR4AP qui représente finement les événements, les acteurs, les relations causales et les dimensions spatio-temporelles afin de rendre les textes directement exploitables. L'article évoque également les choix technologiques et les ressources annotées industrielles. Enfin, nous illustrons comment ces représentations alimentent des interfaces métiers, par exemple pour la veille et la levée d'alertes, démontrant que la structuration sémantique à une granularité fine peut devenir un levier concret de décision industrielle.

ABSTRACT

From a Semantic Representation Formalism to Decision-Making

This paper outlines the main stages of the Prevyo project, from the design of the MR4AP formalism to its implementation in an operational system. It provides an intuition of the usefulness of MR4AP, which finely structures events, actors, causal relationships, and spatio-temporal dimensions to make texts directly actionable. The paper also explains the technological choices as well as some of the industrially annotated resources. Finally, we illustrate how these representations feed business interfaces, for instance in monitoring and alerting applications, demonstrating that fine-grained semantic structuring can become a concrete lever for industrial decision-making.

MOTS-CLÉS : MR4AP, chaîne de traitement, prise de décision.

KEYWORDS: MR4AP, pipeline, decision-making.

1 Introduction

Dans les environnements industriels contemporains, qu'il s'agisse de veille stratégique, de gestion des risques, de cybersécurité ou de sûreté, la prise de décision repose massivement sur l'analyse de flux textuels hétérogènes tels que des dépêches, des rapports, des comptes rendus d'intervention, des publications issues des réseaux sociaux ou encore des documents réglementaires. Le texte brut constitue toutefois une représentation du réel, fondamentalement ambiguë, partielle et dépendante du contexte.

Les approches superficielles, fondées sur des mots-clés, des représentations vectorielles globales ou des classifications thématiques, permettent une première organisation de l'information. Néanmoins,

Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

elles demeurent insuffisantes lorsque l'objectif est de soutenir des décisions critiques. Déclencher une alerte, anticiper une rupture logistique ou évaluer l'impact d'un événement suppose de pouvoir déterminer précisément qui agit, sur quoi, où, quand, selon quelles modalités et avec quelles causes et conséquences.

La structuration fine du texte, au niveau événementiel, argumental et relationnel, devient ainsi un prérequis méthodologique. Elle permet la désambiguïsation des faits, la consolidation d'informations issues de sources multiples, l'inférence temporelle ou causale ainsi que la traçabilité des décisions. Sans représentation sémantique explicite et formalisée, la décision repose sur des signaux fragiles, difficilement interprétables et peu auditables. La question centrale devient alors celle du choix du formalisme de représentation sémantique permettant de passer du texte à la décision.

Dans la section 2, nous rappelerons brièvement le formalisme MR4AP (Giordano & Lopez, 2023) sur lequel s'appuie Prevyo, le système d'IA décrit en section 3. Nous évoquerons certains problèmes liés au peuplement de la base de connaissance en section 4. Nous présenterons en section 5 son utilisation dans la mise à disposition pour l'utilisateur final des informations nécessaires à sa prise de décision.

2 Quel formalisme de représentation sémantique ?

Dans le domaine du traitement automatique des langues, le besoin de formalisations capables de capturer le sens d'un texte de façon exploitable par des machines a donné lieu, au cours des trente dernières années, à une diversité de formalismes. Les approches les plus connues incluent les représentations sous forme d'arbres ou de graphes comme Abstract Meaning Representation (AMR) (Banarescu *et al.*, 2013), qui visent à abstraire l'information d'une phrase sous forme d'un graphe conceptuel, les cadres hérités de la théorie linguistique comme la Discourse Representation Theory (DRT) (Kamp, 1988) ou Minimal Recursion Semantics (MRS) (Copestake *et al.*, 2005), ou encore des formalismes plus cognitifs tels que Universal Conceptual Cognitive Annotation (UCCA) (Abend & Rappoport, 2013). Ces approches se distinguent par leurs objectifs théoriques et le niveau d'abstraction sémantique qu'elles introduisent. Certaines s'abstraient fortement de la syntaxe, d'autres maintiennent des liens étroits avec elle ; certaines préfèrent des représentations continues ou probabilistes, tandis que d'autres reposent sur des labels discrets pour coder les relations.

Malgré leurs avancées fondamentales pour la recherche linguistique et pour des tâches telles que la reconnaissance d'événements, la résolution de coréférences ou encore les paraphrases, ces formalismes présentent des limites lorsqu'on les mobilise directement pour des usages industriels orientés vers la prise de décision opérationnelle. En particulier, beaucoup de ces approches conservent une structuration au niveau de la phrase sans étendre de manière explicite la représentation au contexte documentaire ou multi-sources, ou représentent la sémantique dans des espaces continus rendant l'annotation subjective.

Le formalisme Meaning Representation for Application Purposes [1] (MR4AP) (Giordano & Lopez, 2023) se positionne spécifiquement à l'intersection des exigences linguistiques et des besoins applicatifs industriels. MR4AP est construit autour de cinq principes centraux qui orientent son design vers des représentations explicites, non ambiguës et applicables à une large variété de textes. L'accent est mis sur une visée applicative claire, ce qui signifie que l'annotation doit être unique et non sujette à plusieurs interprétations possibles afin de permettre une exploitation automatisable. Cela réduit le

1. https ://github.com/Emvista/MR4AP

recours à l'expertise linguistique pour résoudre des ambiguïtés d'annotation.

La généricité constitue un autre principe central et garantit que le formalisme est indépendant de la langue et des particularités syntaxiques d'un texte donné. Aux côtés de cette invariance syntaxique, MR4AP mise sur une explicitation maximale des relations et des types d'entités, de manière à ce que chaque relation et chaque acteur soient porteurs de sens clairement interprétable sans nécessiter de glossaire ou d'interprétation contextuelle externe. Cela implique, par exemple, que les rôles sémantiques associés aux arguments d'un événement sont étiquetés au moyen de types suffisamment riches et distincts, inspirés d'inventaires comme VerbNet (Kipper *et al.*, 2006) et sa version française VerbeNet (Danlos *et al.*, 2016), tout en intégrant des catégories spécifiques pour les relations temporelles, spatiales, discursives ou de coréférence.

Une autre dimension essentielle de MR4AP est l'intégration explicite des relations intra- et interphrases, ce qui permet de représenter non seulement les événements isolés mais aussi leur enchaînement, leurs dépendances causales et temporelles à l'échelle d'un document entier. Cette approche dépasse les restrictions de la plupart des formalismes traditionnels qui considèrent souvent chaque phrase comme une unité indépendante ou qui ne modélisent pas systématiquement les relations transphrastiques. Enfin, la richesse des attributs (« attribute richness ») assure que les propriétés associées aux entités, aux événements et aux relations sont suffisamment détaillées pour soutenir des processus d'inférence et de raisonnement automatique.

L'ensemble de ces choix théoriques positionne MR4AP comme un formalisme pragmatique, structuré et orienté vers l'action, ce qui le rend particulièrement adapté à des pipelines de traitement qui visent non seulement à comprendre le texte mais aussi à l'exploiter pour alimenter des bases de connaissance décisionnelles, des moteurs d'alertes ou des systèmes de veille stratégique. MR4AP cherche à la fois à réduire l'ambiguïté, à maximiser l'explicitation des relations sémantiques et à favoriser ainsi l'intégration directe des représentations produites dans des mécanismes automatisés de prise de décision.

3 Du formalisme MR4AP au système d'IA Prevyo

L'adoption d'un formalisme comme MR4AP prend pleinement sens lorsqu'il est opérationnalisé dans un système d'IA capable de structurer automatiquement des flux textuels réels. Le système Prevyo s'inscrit dans cette logique en mettant en œuvre les principes de MR4AP au sein d'une architecture combinant extraction d'événements, reconnaissance d'entités nommées, résolution de coréférences et détection de relations causales et temporelles, entre autres. L'objectif n'est pas seulement d'annoter automatiquement des textes, mais de produire des structures sémantiques directement exploitables dans des environnements décisionnels.

Sur le plan architectural, Prevyo adopte une approche neuro-symbolique qui dépasse les limites des modèles purement neuronaux. Si les architectures Transformer (Vaswani *et al.*, 2017) pré-entraînées sur de larges corpus et adaptées par fine-tuning, offrent des performances remarquables pour la reconnaissance d'entités et la résolution de coréférences, par exemple, elles présentent des fragilités en termes d'explicabilité et de robustesse face à des flux industriels hétérogènes et les performances ne sont plus acceptables dès lors que le graphe sémantique à générer devient complexe (notamment lorsqu'il est régit par un formalisme tel que MR4AP). Prevyo combine donc la puissance de ces modèles pour l'extraction automatique avec des contraintes symboliques alignées sur le formalisme

MR4AP. Cette hybridation améliore la précision des extractions tout en garantissant la cohérence, la traçabilité et la contrôlabilité des informations produites — répondant ainsi aux exigences industrielles de fiabilité et d'auditabilité.

Cette architecture repose sur un écosystème de données annotées constitué au fil de projets collaboratifs. Le projet POPCORN (financement RAPID, Agence Innovation Défense) (Lopez *et al.*, 2024), par exemple, a joué un rôle central en produisant un corpus de rapports de renseignement fictifs et synthétiques en français, annoté en entités nommées, coréférences et relations, et en partie diffusé publiquement [2] (Giordano *et al.*, 2024). Cette même typologie a été appliquée au corpus POPCORN-RENS [3], constitué de transcriptions d'écoutes radio annotées en mentions d'événements (Aubertin *et al.*, 2025). Ces corpus constituent des leviers essentiels pour l'entraînement, l'évaluation et la comparaison de modèles orientés extraction d'information. En complément, Emvista a continué de produire des ressources à l'instar du corpus DWIE-FR [4] annoté en reconnaissance d'entités nommées (Verdy *et al.*, 2023) ou encore un corpus d'e-mails annoté en anaphores (Guenoune *et al.*, 2020) permettant de renforcer les mécanismes intrinsèques de Prevyo tels que la résolution de coréférences. Dans un contexte conversationnel, cela permet à la technologie de consolider sa collecte d'informations répartie à travers plusieurs messages. Les ressources mentionnées, ainsi que celles en cours de création, renforcent la robustesse de Prevyo face à des sources hétérogènes et distribuées.

La maturité du système a été éprouvée depuis 2020 à travers plusieurs preuves de concept industrielles, sur de nombreux cas d'usage, et est actuellement utilisé en production par les clients de Emvista. En collaboration avec la DGA, Prevyo a démontré sa capacité à extraire des informations d'intérêt à destination d'analystes experts (Cousot *et al.*, 2022). Prevyo est également utilisé dans un cadre de simplification documentaire à la SNCF (Reutenauer *et al.*, 2020). Par ailleurs, Prevyo a été appliqué aux e-mails afin de les structurer et extraire automatiquement les tâches (Mekaoui *et al.*, 2020). Ces scénarios d'utilisation qui ont tous mené au peuplement de bases de données ou de connaissance ont peu à peu conduit à l'application centrée utilisateur décrite dans la section suivante.

4 De la base de connaissance aux interfaces utilisateurs

La valeur industrielle de Prevyo se manifeste pleinement lorsque la base de connaissance structurée alimente des interfaces adaptées aux besoins métiers. Dans le cadre de la veille stratégique, les utilisateurs attendent des frises chronologiques interactives, des cartographies dynamiques et des outils de suivi d'acteurs ou de thématiques spécifiques. La structuration événementielle permet alors de naviguer dans l'information selon des axes temporels, géographiques ou relationnels.

À un niveau d'agrégation supérieur, les applications de *Business Intelligence* exploitent les données structurées pour produire des indicateurs agrégés par région, type d'événement ou catégorie d'acteurs. Les tableaux de bord décisionnels et les mécanismes de détection de tendances reposent sur la qualité de la modélisation initiale. Enfin, les applications de levée d'alertes s'appuient sur des règles conditionnelles définies sur des événements formalisés. Le déclenchement automatique d'alertes, leur priorisation selon la gravité ou la proximité, ainsi que leur explicabilité dépendent directement de la richesse de la représentation sémantique.

La structuration conforme à MR4AP offre ici un avantage essentiel : chaque décision peut être

2. https ://github.com/Emvista/popcorn-dataset
3. https ://github.com/Emvista/POPCORN-RENS
4. https ://github.com/Emvista/DWIE-FR

reliée explicitement aux fragments de textes sous-jacents. Ainsi, dans un cas d'usage de détection d'événements, lorsqu'un événement de type ACQUISITION impliquant un acteur stratégique surveillé est détecté dans une ou plusieurs sources concordantes, l'application peut déclencher automatiquement une alerte contextualisée en indiquant l'acteur concerné, la date, la localisation et les sources sous-jacentes. La confiance accordée au système n'est pas aveugle : l'analyste comprend pourquoi une alerte a été levée, sur quelle base et peut en retracer la justification sémantique — garantissant ainsi une traçabilité et une auditabilité compatibles avec les exigences industrielles de responsabilité et de conformité.

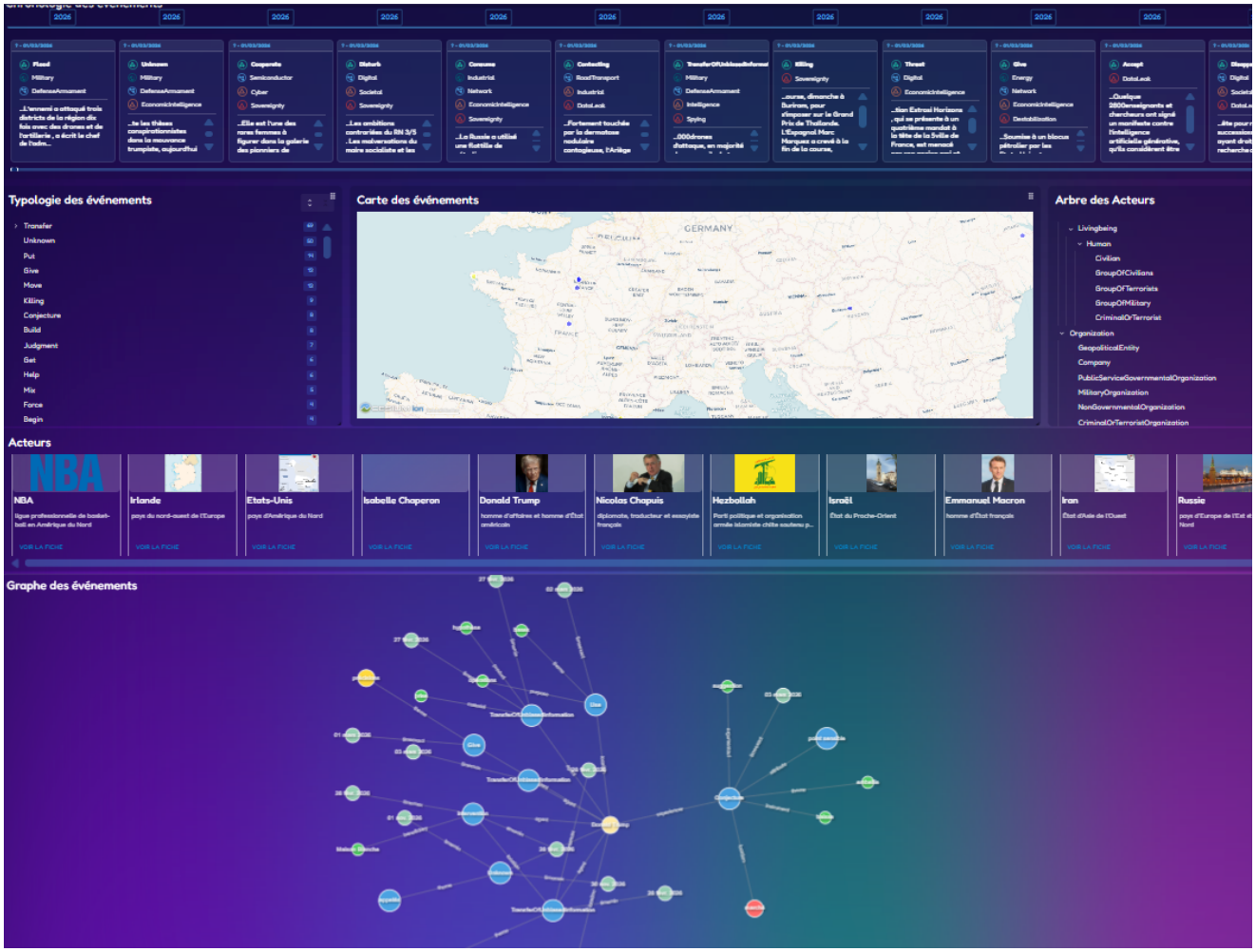

FIGURE 1 – Exemple d'interface utilisateur qui affiche les résultats de Prevyo.

5 Bilan et perspectives

Le passage du texte brut à une décision fiable repose sur une chaîne intégrée allant de la représentation sémantique à l'interface utilisateur, en passant par le système d'IA et la base de connaissance normalisée. Le formalisme MR4AP offre un compromis équilibré entre expressivité linguistique et exploitabilité opérationnelle. Il constitue un pivot structurant permettant l'agrégation multi-source, l'inférence causale et temporelle, ainsi que l'explicabilité des décisions.

Les perspectives de recherche concernent l'amélioration de la robustesse des extractions, l'intégration de connaissances métier dans les processus d'annotation, l'apprentissage conjoint des étapes d'extraction, de fusion et de raisonnement ainsi que l'hybridation entre approches symboliques et neuronales afin de renforcer la fiabilité globale des systèmes. Dans un contexte où les organisations doivent décider plus rapidement sur la base de volumes textuels croissants, la représentation sémantique fine apparaît moins comme un objet académique que comme une infrastructure critique au service de la décision industrielle.

Références

ABEND O. & RAPPOPORT A. (2013). Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 228–238.

AUBERTIN L., GADEK G., SÉRASSET G., PRIEUR M., VUTH N., GRILHERES B., SCHWAB D. & LOPEZ C. (2025). Popcorn-rens : un nouveau jeu de données en français annoté en entités d'intérêts sur une thématique"“sécurité et défense”". In *Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 1–10.

BANARESCU L., BONIAL C., CAI S., GEORGESCU M., GRIFFITT K., HERMJAKOB U., KNIGHT K., KOEHN P., PALMER M. & SCHNEIDER N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, p. 178–186.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

COPESTAKE A., FLICKINGER D., POLLARD C. & SAG I. A. (2005). Minimal recursion semantics : An introduction. *Research on language and computation*, **3**(2), 281–332.

COUSOT K., SANCHEZ T., NGUYEN A., CALPAS A., MARTINEZ G. & LOPEZ C. (2022). Elviria-p : génération d'avis d'expertise pour accompagner les experts en sûreté de fonctionnement des logiciels critiques. *Actes des 33es journées francophones d'Ingénierie des Connaissances*, p. 182.

DANLOS L., PRADET Q., BARQUE L., NAKAMURA T. & CONSTANT M. (2016). Un verbenet du français [a verbnet for french]. *Traitement automatique des langues*, **57**(1), 33–58.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

GIORDANO B. & LOPEZ C. (2023). Mr4ap : Meaning representation for application purposes. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, p. 110–121.

GIORDANO B., PRIEUR M., VUTH N., VERDY S., COUSOT K., SÉRASSET G., GADEK G., SCHWAB D. & LOPEZ C. (2024). Popcorn : Fictional and synthetic intelligence reports for named entity recognition and relation extraction tasks. *Procedia Computer Science*, **246**, 1170–1180.

GUENOUNE H., COUSOT K., LAFOURCADE M., MEKAOUI M. & LOPEZ C. (2020). A dataset for anaphora analysis in french emails. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, p. 165–175.

KAMP H. (1988). Discourse representation theory : What it is and where it ought to go. In *IBM Germany Scientific Symposium Series*, p. 84–111 : Springer.

KIPPER K., KORHONEN A., RYANT N. & PALMER M. (2006). Extending verbnet with novel verb classes. In *LREC*, p. 1027–1032 : Genoa.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LOPEZ C., VERDY S., GADEK G., PRIEUR M., SCHWAB D., SÉRASSET G. & VUTH N. (2024). Popcorn : Ia d'extraction d'information à partir de sources textuelles pour le renseignement militaire. In *6th Conference on Artificial Intelligence for Defense*.

MEKAOUI M., TISSERANT G., DODARD M. & LOPEZ C. (2020). Extraction de tâches dans les e-mails : une approche fondée sur les rôles sémantiques. In *EGC*, p. 193–204.

REUTENAUER C., LEFEUVRE L., FOUQUERAY A., PROUTEAU T., PELLOIN V., LOPEZ C., NATHALIE C., SEGOND F., NICOLAS D. & BOURIGAULT D. (2020). Technologies sémantiques et accès à l'information dans le prescrit sncf. In *Congrès Lambda Mu 22 «Les risques au cœur des transitions»(e-congrès)-22e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Institut pour la Maîtrise des Risques*.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VERDY S., PRIEUR M., GADEK G. & LOPEZ C. (2023). Dwie-fr : Un nouveau jeu de données en français annoté en entités nommées. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux–articles courts*, p. 63–72.