ding-01 :ARG0
un corpus AMR pour le français parlé spontané

Jeongwoo Kang[1] Maria Boritchev[2] Maximin Coavoux[1]
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2] Paris Télécom, 19 place Marguerite Perey, 91120 Palaiseau, France
{jeongwoo.kang, maximin.coavoux}@univ-grenoble-alpes.fr
maria.boritchev@telecom-paris.fr

RÉSUMÉ

Nous présentons notre travail en cours sur l'annotation d'un corpus sémantique du français. Nous annotons le corpus DinG, constitué de transcriptions de dialogues spontanés en français enregistrées pendant des parties du jeu de plateau *Catan*, en *Abstract Meaning Representation* (AMR), un formalisme de représentation sémantique. Comme AMR a une couverture insuffisante de la dynamique de la parole spontanée, nous étendons le formalisme pour mieux représenter la parole spontanée et les structures de phrases spécifiques au français. En outre, nous diffusons un guide d'annotation détaillant ces extensions. Enfin, nous publions notre corpus sous licence libre (CC-SA-BY). Notre travail contribue au développement de ressources sémantiques pour le dialogue en français.

ABSTRACT

ding-01 :ARG0 an AMR corpus for spontaneous spoken French

We present our ongoing work on the annotation of a semantic corpus in French. We annotate the corpus DinG of transcriptions of spontaneous dialogues between players of *Catan*, in Abstract Meaning Representation (AMR), one of the most popular meaning representation frameworks. Since AMR does not cover some features of dialogue dynamics, we extend the framework to better represent spoken language as well as sentence structures specific to French. In addition, we provide an annotation guide detailing these extensions. Finally, we publish our corpus under a Creative Commons license (CC-SA-BY). Our work contributes to the development of semantic resources for French dialogue.

MOTS-CLÉS : Annotation, sémantique, AMR, parole, corpus.

KEYWORDS: Annotation, semantics, AMR, speech, corpus.

1 Introduction

Abstract Meaning Representation (Banarescu *et al.*, 2013, AMR) code le sens d'un texte sous la forme d'un graphe enraciné, dirigé et acyclique (voir figure 1). La représentation du sens sous une forme structurée présente plusieurs avantages pour les systèmes d'information par rapport au langage naturel. Les représentations AMR réduisent l'ambiguïté sémantique en spécifiant explicitement une interprétation plausible parmi d'autres. En outre, comme AMR fait abstraction des variations de surface –notamment des variations syntaxiques–, les phrases ayant le même sens sous-jacent partagent la même représentation AMR (*e.g.*, « La police a arrêté le voleur. » et « Le voleur a été arrêté par

la police. »). Cette représentation canonique permet de réduire l'espace de recherche des modèles, ce qui fait d'AMR un outil utile pour diverses tâches de TAL, telles que la traduction automatique (Wein & Schneider, 2024), le résumé automatique de texte (Liao *et al.*, 2018; Liu *et al.*, 2015) et l'interaction humain·e-robot (Bonial *et al.*, 2019, 2023).

L'entraînement d'un analyseur AMR pour générer automatiquement un graphe AMR à partir d'un texte donné nécessite un ensemble de données composé de textes associés à leurs graphes AMR correspondants. Cependant, les ensembles de données AMR pour le français sont actuellement rares, puisque la plupart des ressources AMR disponibles sont pour l'anglais. Ce déséquilibre dans les ressources sémantiques limite le développement d'analyseurs sémantiques pour le français, ce qui entrave le progrès des systèmes TAL qui en dépendent. En outre, la plupart des données AMR existantes sont basées sur des textes écrits tels que des articles de presse et des forums en ligne. En revanche, les données de dialogue, qui présentent des caractéristiques linguistiques uniques en raison de leur nature interactive et spontanée (*e.g.*, les marqueurs de discours comme *alors, du coup, donc*, ou les *backchannel* [1], ou encore les disfluences : faux départs, hésitations) restent sous-représentées.

Pour combler cette lacune dans les ressources sémantiques françaises, en particulier pour le dialogue, nous annotons manuellement le corpus DinG (Boritchev & Amblard, 2022) en AMR. DinG est constitué de transcriptions de dialogues enregistrés lors de sessions du jeu de plateau *Catan*, capturant diverses caractéristiques linguistiques de l'interaction orale en français. Cependant, le formalisme AMR standard, tel qu'il est défini actuellement [2], présente des limites dans la représentation des caractéristiques spécifiques à la parole. Par conséquent, nous étendons AMR en introduisant des relations supplémentaires pour (i) annoter deux phénomènes pragmatiques : les marqueurs de discours et les expressions de *backchannel*, (ii) représenter la coréférence à travers plusieurs tours de parole.

Pour résumer, nos contributions principales sont les suivantes [3] :

— Nous publions `ding-01` [4], un nouveau corpus AMR du français parlé spontané de 1169 tours de parole. Le corpus a vocation à être augmenté et à couvrir 3000 tours de parole d'ici fin 2025. Nous diffusons également un *data statement* avec le corpus pour décrire l'ensemble des métadonnées pertinentes et biais potentiels, suivant les bonnes pratiques de production et de diffusion de données pour le TAL (Bender & Friedman, 2018; McMillan-Major *et al.*, 2024).
— Nous adaptons AMR pour représenter les phénomènes de parole spontanée en français, y compris les marqueurs de discours et les *backchannels*.
— Nous fournissons un guide d'annotation pour 1) assurer la cohérence des annotations en clarifiant les aspects non spécifiés dans le guide d'annotation original AMR, 2) définir comment annoter les caractéristiques linguistiques spécifiques au français.

1. Dans un dialogue, le *backchannelling* comprend les indices, verbaux ou non, qui indiquent à l'interlocuteurice que l'on prête attention à ce qu'iel dit : hochements de tête, *ok, hum*, etc.

2. La version actuelle du guide d'annotation est disponible sur https://github.com/amrisi/amr-guidelines/blob/master/amr.md

3. Les ressources que nous publions (corpus, guide, *data statement*) sont disponibles sur Zenodo : https://doi.org/10.5281/zenodo.15537426

4. https://propbank.github.io/v3.4.0/frames/ding.html#ding.01

2 Contexte et travaux connexes

2.1 Introduction à AMR

AMR représente le sens des textes par des graphes orientés acycliques et enracinés. Dans un graphe AMR, les nœuds sont :

— soit des prédicats prédéfinis dans Propbank [5] (Palmer *et al.*, 2005), *e.g.*, `break-01` dans la figure 1,
— soit des noms en anglais, *e.g.*, `man` et `window` dans la figure 1,
— soit des mots-clés spécifiques à AMR, *e.g.*, `date-entity`.

Les arcs du graphe AMR sont étiquetées pour indiquer la relation entre les nœuds. Par exemple, `:ARG0` et `:ARG1` dans la figure 1 indiquent respectivement que `man` est l'agent du prédicat `break-01` et que `window` est l'objet de ce même prédicat, conformément à la structure argumentale décrite dans Propbank [6]. Un graphe AMR peut aussi être représenté sous forme textuelle (voir figure 2). Bien qu'AMR ait été initialement conçue pour les textes en anglais, elle est couramment utilisée pour représenter des textes autres qu'en anglais (Damonte & Cohen, 2018; Xu *et al.*, 2021; Liu *et al.*, 2020). Deux phrases qui expriment le même chose dans deux langues différentes (deux phrases en relation de traduction) partagent un même graphe AMR.

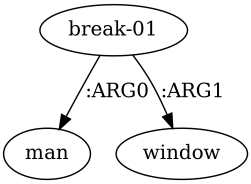

FIGURE 1 – Graphe AMR pour "A man breaks a window" ou « Un homme a cassé la fenêtre ».

```
(b / break-01
    :ARG0 (m / man)
    :ARG1 (w / window))
```

FIGURE 2 – Graphe AMR linéarisé dans un format texte.

2.2 Les jeux de données d'AMR

La plupart des données d'AMR à grande échelle, y compris AMR 3.0 (Knight *et al.*, 2020) et Massive-AMR (Regan *et al.*, 2024), sont disponibles exclusivement en anglais. AMR 3.0 est le jeu de données le plus populaire pour l'entraînement et l'évaluation des analyseurs d'AMR. Il contient environ 60 000 exemples de textes provenant de diverses sources telles que des articles d'actualité, des blogs et des forums en ligne. Massive-AMR, le plus grand jeu de données AMR annoté manuellement, se compose de 84 000 énoncés adressés à un assistant virtuel, chacun associé à son graphe AMR correspondant. La plupart des phrases dans Massive-AMR sont de brèves questions ou demandes.

Pour le français, quelques jeux de données sont disponibles : *Le Petit Prince* AMR (Kang *et al.*, 2023), Massive-AMR français (Regan *et al.*, 2024) et ReMEDIATE (Druart, 2024). Pour *Le petit Prince* AMR, les auteurs ont aligné manuellement *The Little Prince* AMR [7] anglais avec le texte

5. https://propbank.github.io/v3.4.0/frames/
6. https://propbank.github.io/v3.4.0/frames/break.html#break.01
7. https://github.com/flipz357/AMR-World/blob/main/data/reference_amrs/amr-bank-struct-v3.0.txt

originel en français, tandis que le Massive-AMR français est constitué d'une partie de Massive-AMR anglais (Regan *et al.*, 2024), manuellement traduite en français. ReMEDIATE est annoté de façon semi-automatique en français, sans recourir à des données préexistantes en anglais. En terme de type des corpus, *Le Petit Prince* AMR est une œuvre littéraire. Massive-AMR se compose de requêtes envoyées à des assistants virtuels. Enfin, ReMEDIATE contient des interactions entre un assistant virtuel pour effectuer des réservations et son utilisateurice. Notons que ce corpus reprend la syntaxe des graphes AMR mais adapte l'ensemble des concepts et étiquettes d'arcs utilisés à la tâche de compréhension de la parole (*spoken language understanding*).

Notre travail se distingue sur plusieurs points essentiels. Premièrement, nous annotons des conversations spontanées entre plusieurs locuteurices. Notre corpus capture des interactions réelles, reflétant la dynamique de la parole spontanée en français. De plus, *Le Petit Prince* AMR et Massive-AMR ont été initialement annotés en anglais, puis adaptés à d'autres langues par la traduction manuelle ou l'alignement translingue (en exploitant l'hypothèse que des phrases en relation de traduction doivent avoir le même graphe sémantique). Ce processus peut introduire des biais, rendant les données possiblement centrées sur l'anglais. Nous annotons directement des dialogues français en AMR sans dépendre des annotations antérieures en anglais, en s'assurant que la sémantique du français soit préservée tout au long du processus d'annotation. Enfin, alors que REMEDIATE est annoté de manière semi-automatique, nous annotons les données manuellement. Il est important de préciser ici que l'annotation sémantique n'est pour le moment pas réalisée de manière satisfaisante par de grands modèles de langues, y compris pour l'anglais (Ettinger *et al.*, 2023).

2.3 AMR pour des dialogues

Même si AMR standard fournit divers rôles sémantiques pour présenter des sens des textes, plusieurs efforts ont été faits pour l'étendre afin de capturer divers aspects du dialogue. DMR (Hu *et al.*, 2022) et Dialogue-AMR (Bonial *et al.*, 2020), ainsi que le travail de Druart (2024) font partie de ces extensions. Ces trois approches se concentrent principalement sur les dialogues axés sur des tâches, dans lesquels un·e agent·e demande une action à un agent robotique ou virtuel. Par conséquent, elles intègrent des instructions fines et introduisent des rôles supplémentaires pour représenter, par exemple, la force illocutoire ou la contribution intentionnelle des locuteurices (Bonial *et al.*, 2020).

Cependant, ces rôles ne conviennent pas parfaitement à notre corpus, composé de conversations spontanées entre plusieurs interlocuteurices. Nous nous efforçons d'adhérer le plus possible aux conventions d'AMR standard, en suivant les guides d'annotations définis actuellement. Néanmoins, notre corpus étant en français et issu de la parole spontanée, il présente des caractéristiques linguistiques spécifiques aux interactions orales naturelles, telles que les *backchannels* et les marqueurs de discours.

Les marqueurs de discours et les *backchannels* véhiculent des informations pragmatiques dans le dialogue. Cependant, l'AMR standard ne prend pas en compte ce type d'information, comme cela est précisé dans son guide d'annotation. Malgré cela, nous avons choisi d'annoter les informations pragmatiques transmises par les marqueurs discursifs pour deux raisons principales. Premièrement, contrairement à AMR 3.0 qui repose principalement sur des données textuelles, notre corpus est constitué de dialogues riches en contenus pragmatiques. Nous estimons que l'annotation de ces informations offre une ressource précieuse pour l'étude du dialogue en français. De plus, les rôles supplémentaires que nous proposons peuvent être facilement supprimées, ce qui garantit la compatibilité avec AMR 3.0.

Deuxièmement, bien que le guide d'annotation d'AMR stipule que l'information pragmatique n'est pas incluse, en pratique, AMR intègre certains aspects pragmatiques. Par exemple, le choix du nœud racine dans AMR dépend souvent de l'élément sur lequel porte l'attention principale de la phrase, ce qui reflète une information pragmatique. Ainsi, l'ajout d'éléments pragmatiques à nos annotations n'est pas totalement incompatible avec les pratiques d'AMR standard. Pour tenir compte de ces informations pragmatiques, nous introduisons de nouveaux rôles, qui sont détaillés dans la section 5.

3 Le corpus DinG

Nous annotons le corpus DinG [8] (Boritchev & Amblard, 2022), une collection de dialogues multipartites transcrits manuellement entre des joueureuses francophones du jeu de société Catan [9]. Catan est un jeu de plateau stratégique centré sur la gestion et l'échange des ressources. Ainsi les joueureuses négocient souvent des échanges de ressources entre elleux et leurs interactions réelles sont enregistrées dans le corpus. Nous avons sélectionné ce corpus pour deux raisons essentielles.

Premièrement, DinG est disponible sous une licence libre [10]. Notre objectif étant de rendre nos données publiques, la sélection de données sources sous licence libre était une exigence cruciale. Deuxièmement, DinG est constitué de dialogues naturels entre locuteurices. Comme l'environnement n'était pas contrôlé par les collecteurices de données et que les joueureuses étaient libres d'interagir pendant le jeu, ce jeu de données capture un flux conversationnel naturel et comprend une grande variété de phénomènes dialogiques. Le corpus sémantique constituera donc un ensemble de test idéal pour l'évaluation des modèles de langue préentraînés sur le domaine des transcriptions de parole spontanée.

4 DinG-AMR

Dans cette section, nous présentons quelques statistiques sur le corpus, le processus d'annotation et la qualité des données évaluée par accord inter-annotateurices. L'annotation s'est déroulée sur une période de trois mois. Au total, nous avons annoté 1169 tours de paroles en AMR [11].

Parmi ces exemples, on dénombre 310 marqueurs de discours et 19 instances de *backchannel*. L'annotation du corpus a été principalement réalisée par la première autrice de cet article à l'aide de l'outil d'annotation *metAMoRphosED* (Heinecke, 2023). Environ 15 % des exemples du corpus entier ont été validés par deux autres annotateurices, coauteurices de cet article. Plus précisément, l'annotatrice principale et les deux annotateurices se sont réunies régulièrement tout au long du processus d'annotation (une fois par semaine ou toutes les deux semaines) afin de vérifier la validité des exemples un par un et répertorier les difficultés rencontrées. En cas de désaccord entre les trois annotateurices, l'exemple était corrigé ou modifié au cours de la discussion.

Nous avons rencontré plusieurs difficultés lors du processus d'annotation (le processus d'annotation est illustré en figure 3). L'un des exemples concernait le mot '*donc*', qui apparaît fréquemment dans DinG. Dans la majorité des cas, il fonctionne davantage comme un marqueur discursif que

8. https://gitlab.inria.fr/semagramme-public-projects/resources/ding/
9. Nous renvoyons les lecteurices vers le site https://www.catan.com/ pour plus d'information sur le jeu.
10. La licence *Attribution ShareAlike Creative Commons* (CC BY-SA 4.0).
11. Nous avons gardé le découpage en tours de parole fait dans le corpus DinG.

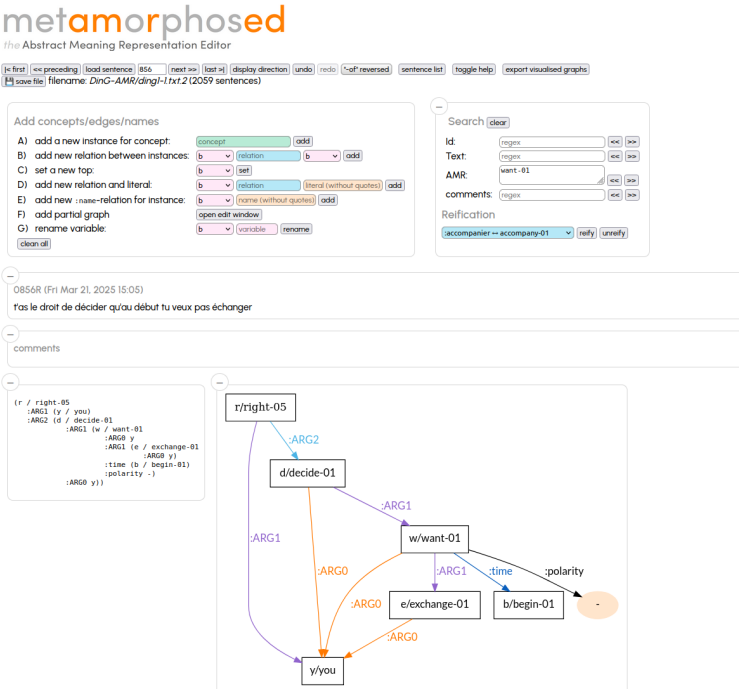

FIGURE 3 – Capture d'écran du processus d'annotation des données

comme une conjonction causale. Toutefois, son usage était souvent ambigu, et les deux interprétations pouvaient être valides selon le contexte. Afin de réduire l'ambiguïté et d'améliorer la cohérence entre les annotations, nous avons établi la règle suivante : annoter '*donc*' systématiquement comme un marqueur discursif, à condition que sa suppression ne modifie pas le sens de la phrase. Notre méthode pour traiter d'autres difficultés similaires en définissant des lignes directrices claires est détaillée dans notre guide d'annotation. De plus, lorsque nous étions confrontés à des cas complexes, ou des cas où plusieurs choix d'annotations étaient corrects, nous nous sommes systématiquement référés aux données existantes de l'AMR 3.0 en anglais pour choisir entre les différentes possibilités. Les tours de paroles concernés contiennent en commentaires les références des phrases AMR 3.0 qui justifient ces choix.

Afin d'évaluer la qualité des annotations, 160 exemples issus de notre corpus ont été annotés par deux annotateurices. L'accord entre elleux a été mesuré à l'aide du score SMATCH (Cai & Knight, 2013), une mesure d'évaluation pour AMR calculée en comptant le nombre de triplets (concept, arc étiqueté, concept) en commun. Nous avons obtenu un score de 71,3. À titre de comparaison, Banarescu *et al.* (2013) rapportent des scores d'accord inter-annotateurices allant de 71 à 83, en fonction de la source des données et du niveau d'expertise des annotateurices.

À la suite de cette évaluation, nous avons effectué une étape de résolution de conflits d'annotation, afin de produire notre corpus *gold* final. Cette étape est une vérification par les trois auteurices de cet article, ensemble, des 160 exemples doublement annotés. Pour les exemples présentant des désaccords entre les deux annotateurices [12], un choix a été fait, soit pour une des deux annotations proposées, soit

12. Un total de 106 exemples présentaient des désaccords

pour une troisième annotation décidée à trois.

Des conflits fréquents portaient sur les étiquettes d'arcs (`:ARG0`, `:ARG1`, `:ARG2`) relevant d'erreurs d'annotations facilement corrigées une fois détectées. Un autre type de conflit concerne le choix de concepts propbank synonymes. Par exemple, `own-01` [13] et `possess-01` [14] représentent le même concept et ont les deux mêmes rôles sémantiques (`:ARG0` *possédant*, `:ARG1` *possédé*). Dans les données AMR de l'anglais, le choix entre ces deux concepts dépend de l'item lexical utilisé dans la phrase. Nous avons utilisé ce type de conflits pour affiner le guide d'annotation de manière à toujours faire le même choix entre les deux concepts (dans ce cas : `own-01`).

5 AMR adapté au corpus DinG

Tout en adhérant aussi étroitement que possible à AMR standard, nous introduisons quelques extensions pour mieux représenter les caractéristiques de la parole spontanée française. Certaines de ces caractéristiques clés sont décrites ci-dessous. En outre, nous annotons la coréférence inter-instances dans le corpus, ce qui le distingue d'AMR 3.0. Nous adaptons également le concept standard AMR de *focus* pour présenter une structure de phrase typique au français. Des informations détaillées sur ces annotations peuvent être trouvées dans notre guide d'annotation.

Pour assurer une compatibilité avec le corpus anglais AMR 3.0 pour les cas d'usages de DinG-AMR qui le nécessiteraient, ces extensions peuvent être facilement supprimées des graphes AMR.

5.1 Marqueurs de discours

Les marqueurs de discours sont de courts mots ou phrases utilisés par les interlocuteurices pour structurer leur discours (*e.g.*, *donc*, *et*). Ils sont utilisés pour commencer un tour de parole, ou peuvent servir de *filler* au milieu d'un énoncé, pendant une hésitation. Nous introduisons un nouveau rôle, `:discourse-marker`, pour les annoter (voir figure 4). Ce nouveau rôle est réifiable avec le concept `be-discourse-marker-91`.

5.2 Backchannel

Les *backchannel* se réfèrent à de courtes interjections faites par un·e participant·e à la conversation pendant qu'un·e autre locuteurice parle (e.g., *hum*, *mmh-mmh*) pour manifester son attention à la conversation. Nous les annotons en utilisant une nouvelle relation `:back-channel` réifiable avec le concept `be-back-channel-91` (voir figure 5).

13. https://propbank.github.io/v3.4.0/frames/own.html#own.01
14. https://propbank.github.io/v3.4.0/frames/possess.html#possess.01

```
(p / put-01
   :ARG0 (y / you)
   :ARG1 (r / road)
   :mode imperative
   :ARG2 (h / here)
   :polarity -
   :discourse-marker "donc")
```

FIGURE 4 – Graphe AMR pour « **Donc** mets pas ta route ici ».

```
(b / be-back-channel-91
   :ARG2 "hum")
```

FIGURE 5 – Graphe AMR pour « hum ».

5.3 Co-référence inter-instances

Puisque le corpus DinG capture les interactions entre les joueureuses tout au long du jeu, la coréférence peut s'étendre sur plusieurs énoncés ou instances. Pour assurer une représentation complète du sens, nous annotons les coréférences multi-instances, en marquant les antécédents qui apparaissent dans différents énoncés. Le nœud `s0080b_s_stone` de la figure 7 indique que son antécédent provient de l'exemple identifié par l'ID `0080b` de la figure 6 et du concept `s / stone` associé à cet exemple.

```
# ::id 0080b
(w / want-01
   :ARG0 (y / you)
   :ARG1 (s / stone)
   :polarity (a / amr-unknown))
```

FIGURE 6 – « Tu veux de la **pierre** ? »

```
# ::id 0082b
(e / exchange-01
   :ARG0 (I / I)
   :ARG2 (y / you)
   :ARG1 (s / sheep
      :quant 3)
   :ARG3 (s1 / s0080b_s_stone))
```

FIGURE 7 – « Je te **l**'échange contre 3 moutons »

5.4 Représentations du focus

En AMR, le *focus* d'une phrase est indiqué par la racine du graphe. Nous appliquons ce principe à l'annotation de la structure clivée, une structure de phrase couramment utilisée en français pour la focalisation. La structure clivée est une construction suivant le patron suivant « c'est [sujet] qui... » servant à mettre en valeur le [sujet]. Afin de refléter cet accent mis sur le sujet, nous le sélectionnons comme racine du graphe AMR. La figure 8 présente un exemple de phrase construite de manière clivée, accompagnée de son annotation en AMR. Nous adoptons la même stratégie pour les cas de dislocations à gauche avec reprise pronominale, comme dans l'énoncé : « moi je veux 2 blés ». Ce type de structures, très courant à l'oral, est également une façon d'exprimer un focus en français. Dans ce cas, le concept `i` sera la racine du graphe AMR.

5.5 Disfluences

Les disfluences sont fréquentes dans les dialogues spontanés. On observe souvent des marqueurs de disfluence (*e.g.*, *euh*, *eh*), des répétitions (*e.g.*, *franchement t'es t'es franchement*) et des faux départs (*e.g.*, *j'ai be- j'ai pas de bois*) dans le corpus DinG. Dans AMR standard, les marqueurs de disfluences ne sont pas annotés. Conformément à cette convention, nous n'annotons pas les marqueurs

de disfluences, les répétitions ou les faux départs courts. Cependant, si un faux départ a un contenu sémantique interprétable, nous l'annotons en utilisant `multi-sentence` (voir figure 9).

```
(y / you
   :ARG0-of (c / choose-01
      :ARG1 (p1 / place
         :ARG2-of (p / put-01
            :ARG1 (t / they))))
   :polarity (a / amr-unknown))
```

FIGURE 8 – « **C'est toi qui** choisis où est-ce que tu les mets ? »

```
(m / multi-sentence
  :snt1 (d / do-02
        :ARG0 (I / I))
  :snt2 (t / that
        :ARG2-of (o / obligate-01
              :purpose (m1 / make-01
                    :ARG1 (s / settlement)
                    :location (h / here))
              :ARG1-of (r /
request-confirmation-91))))
```

FIGURE 9 – « **Je vais faire une** euh c'est ça hein qu'il faut pour faire une une colonie ici »

6 Conclusion et travaux futurs

Nous avons présenté nos travaux en cours pour annoter le corpus DinG en AMR afin d'enrichir les données d'annotation sémantique en français. Pour mieux représenter la dynamique de la parole spontanée dans le corpus DinG, nous avons adapté le standard AMR en introduisant de nouvelles relations. Nous mettons à disposition un guide d'annotation détaillant ces adaptations, ainsi qu'une déclaration de données (*data statement*) contenant les métadonnées de DinG-AMR. Dans nos travaux futurs, nous prévoyons d'annoter les ellipses présentes dans le corpus, en particulier les ellipses verbales, qui se produisent fréquemment lorsque le verbe d'une phrase n'est pas répété d'un tour de parole à l'autre, mais reste facilement inférable. Pour assurer une représentation complète des énoncés, nous intégrerons ces ellipses dans l'annotation, et la version mise à jour du corpus en tiendra compte. Nous avons pour objectif d'étendre l'ensemble de données annotées à 3 000 tours de parole.

Remerciements

Nous remercions les trois relecteurices anonymes pour leurs nombreuses remarques. Ce travail a bénéficié de financements de l'institut Carnot Cognition (projet ANAGRAM) et de l'Agence Nationale de la Recherche, via le projet SynPaX (ANR-23-CE23-0017-01).

Références

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 A. PAREJA-LORA, M. LIAKATA & S. DIPPER, Éds., *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 178–186, Sofia, Bulgaria : Association for Computational Linguistics.

BENDER E. M. & FRIEDMAN B. (2018). Data statements for natural language processing : Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, **6**, 587–604. DOI : 10.1162/tacl_a_00041.

BONIAL C., DONATELLI L., ABRAMS M., LUKIN S. M., TRATZ S., MARGE M., ARTSTEIN R., TRAUM D. & VOSS C. (2020). Dialogue-AMR : Abstract Meaning Representation for dialogue. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 684–695, Marseille, France : European Language Resources Association.

BONIAL C., FORESTA J., FUNG N. C., HAYES C. J., OSTEEN P., ARKIN J., HEDEGAARD B. & HOWARD T. (2023). Abstract Meaning Representation for grounded human-robot communication. In J. BONN & N. XUE, Éds., *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, p. 34–44, Nancy, France : Association for Computational Linguistics.

BONIAL C. N., DONATELLI L., ERVIN J. & VOSS C. R. (2019). Abstract Meaning Representation for human-robot dialogue. In G. JAROSZ, M. NELSON, B. O'CONNOR & J. PATER, Éds., *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, p. 236–246. DOI : 10.7275/v3c5-yd35.

BORITCHEV M. & AMBLARD M. (2022). A multi-party dialogue ressource in French. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 814–823, Marseille, France : European Language Resources Association.

CAI S. & KNIGHT K. (2013). Smatch : an evaluation metric for semantic feature structures. In H. SCHUETZE, P. FUNG & M. POESIO, Éds., *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 748–752, Sofia, Bulgaria : Association for Computational Linguistics.

DAMONTE M. & COHEN S. B. (2018). Cross-lingual Abstract Meaning Representation parsing. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1146–1155, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1104.

DRUART L. (2024). *Vers une Compréhension Contextuelle et Structurée de la Parole Dialogique Orientée Tâche*. Theses, Université d'Avignon. HAL : tel-04963857.

ETTINGER A., HWANG J., PYATKIN V., BHAGAVATULA C. & CHOI Y. (2023). "you are an expert linguistic annotator" : Limits of LLMs as analyzers of Abstract Meaning Representation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 8250–8263, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.553.

HEINECKE J. (2023). metAMoRphosED, a graphical editor for Abstract Meaning Representation. In H. BUNT, Éd., *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, p. 27–32, Nancy, France : Association for Computational Linguistics.

HU X., DAI J., YAN H., ZHANG Y., GUO Q., QIU X. & ZHANG Z. (2022). Dialogue meaning representation for task-oriented dialogue systems. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2022*, p. 223–237, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-emnlp.17.

KANG J., COAVOUX M., SCHWAB D. & LOPEZ C. (2023). Analyse sémantique AMR pour le français par transfert translingue. In C. SERVAN & A. VILNAT, Éds., *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. 55–62, Paris, France : ATALA.

KNIGHT K., BADARAU B., BARANESCU L., BONIAL C., BARDOCZ M., GRIFFITT K., HERMJAKOB U., MARCU D., PALMER M., O'GORMAN T. & SCHNEIDER N. (2020). Abstract meaning representation (amr) annotation release 3.0 ldc2020t02. Philadelphia : Linguistic Data Consortium. DOI : https ://doi.org/10.35111/44cy-bp51.

LIAO K., LEBANOFF L. & LIU F. (2018). Abstract Meaning Representation for multi-document summarization. In E. M. BENDER, L. DERCZYNSKI & P. ISABELLE, Éds., *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1178–1190, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

LIU F., FLANIGAN J., THOMSON S., SADEH N. & SMITH N. A. (2015). Toward abstractive summarization using semantic representations. In R. MIHALCEA, J. CHAI & A. SARKAR, Éds., *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1077–1086, Denver, Colorado : Association for Computational Linguistics. DOI : 10.3115/v1/N15-1114.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

MCMILLAN-MAJOR A., BENDER E. M. & FRIEDMAN B. (2024). Data statements : From technical concept to community practice. *ACM J. Responsib. Comput.*, **1**(1). DOI : 10.1145/3594737.

PALMER M., GILDEA D. & KINGSBURY P. (2005). The Proposition Bank : An annotated corpus of semantic roles. *Computational Linguistics*, **31**(1), 71–106. DOI : 10.1162/0891201053630264.

REGAN M., WEIN S., BAKER G. & MONTI E. (2024). MASSIVE multilingual Abstract Meaning Representation : A dataset and baselines for hallucination detection. In D. BOLLEGALA & V. SHWARTZ, Éds., *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, p. 1–17, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.starsem-1.1.

WEIN S. & SCHNEIDER N. (2024). Lost in translationese ? reducing translation effect using Abstract Meaning Representation. In Y. GRAHAM & M. PURVER, Éds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 753–765, St. Julian's, Malta : Association for Computational Linguistics.

XU D., LI J., ZHU M., ZHANG M. & ZHOU G. (2021). XLPT-AMR : Cross-lingual pre-training via multi-task learning for zero-shot AMR parsing and text generation. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 896–907, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.73.