



Titre: Extraction des connaissances en géométrie plane à partir d'énoncés
Title: de problèmes

Auteur: Othmane Farid
Author:

Date: 2020

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Farid, O. (2020). Extraction des connaissances en géométrie plane à partir
Citation: d'énoncés de problèmes [Master's thesis, Polytechnique Montréal]. PolyPublie.
<https://publications.polymtl.ca/5374/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/5374/>
PolyPublie URL:

Directeurs de recherche: Michel Gagnon, Fabienne Venant, & Philippe R. Richard
Advisors:

Programme: Génie informatique
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Extraction des connaissances en géométrie plane à partir d'énoncés de problèmes

OTHMANE FARID

Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie informatique

Août 2020

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Extraction des connaissances en géométrie plane à partir d'énoncés de problèmes

présenté par **Othmane FARID**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Michel DAGENAIS, président

Michel GAGNON, membre et directeur de recherche

Fabienne VENANT, membre et codirectrice de recherche

Philippe R. RICHARD, membre et codirecteur de recherche

Michel DESMARAIS, membre

REMERCIEMENTS

Je voudrais dans un premier temps remercier mes directeurs de recherche, Michel Gagnon, Fabienne Venant et Philippe R. Richard, de m'avoir encadré, orienté, conseillé et financé. Je remercie particulièrement Michel Gagnon et Fabienne Venant pour m'avoir énormément aidé à bien structurer et corriger ce mémoire. Je remercie également l'Ecole Nationale d'Informatique et de mathématiques Appliquées de Grenoble pour m'avoir permis de mener à bien ce projet de double-diplôme en partenariat avec l'Ecole Polytechnique de Montréal. Enfin, je remercie mes parents et Isabel Almario qui ont toujours été là pour moi. Leurs soutiens inconditionnels et leurs encouragements ont été d'une grande aide.

RÉSUMÉ

L'émergence actuelle des tuteurs intelligents est appelée à transformer les modes d'enseignement traditionnels. Ces tuteurs intelligents permettent un accompagnement et un soutien personnalisés à chaque élève, et ainsi diminuent la surcharge de travail pour l'enseignant, qui peut alors consacrer plus de temps aux élèves qui ont besoin d'un soutien particulier. Dans ce travail, nous nous intéressons aux tuteurs intelligents qui permettent d'accompagner et d'aider les étudiants de niveau secondaire à résoudre des questions de la géométrie plane. Malheureusement, dans les tuteurs intelligents actuels, et en particulier dans les tuteurs destinés aux utilisateurs qui parlent français, l'extraction automatique des connaissances est laissée aux auteurs du problème. Ces derniers sont amenés à extraire les informations du problème et à les interpréter avant de les saisir manuellement dans l'interface utilisateur qui est sous forme d'une structure d'entrée prédéfinie dans le tuteur (par exemple, une liste des hypothèses et des conclusions). Ce type de tuteur risque de donner des résultats erronés à la suite d'une mauvaise ou une incomplète interprétation. De ce constat naît la motivation de notre recherche, qui est la création d'un extracteur automatique des connaissances à partir d'énoncés de problèmes de géométrie plane écrits en français. Cet extracteur automatisera l'ajout de nouveaux problèmes dans le tuteur. Cet extracteur s'inscrit dans le projet QED-Tutrix, qui a pour but de créer un tuteur intelligent pour la géométrie plane telle qu'enseignée dans le contexte scolaire québécois au niveau secondaire.

L'extracteur que nous proposons est basé sur l'analyse syntaxique et grammaticale de l'énoncé, qui est combinée à un ensemble de règles pour extraire les connaissances. L'approche adoptée consiste à adapter à la langue française un extracteur déjà existant qui a été conçu pour traiter les questions à choix multiples écrites en anglais. Cette adaptation a été évaluée sur les questions traduites en français et puis sur un ensemble de questions tirées des manuels scolaires québécois. Notre approche, qui se fonde sur une meilleure représentation formelle de la sortie de l'extracteur (c'est-à-dire l'encodage des relations extraites de la question) donne de bons résultats, non seulement pour les questions à choix multiple, mais aussi pour les questions inspirées des manuels scolaires québécois.

ABSTRACT

The current emergence of smart tutoring systems is called upon to transform traditional teaching methods. These intelligent tutors allow personalized guidance and support for each student, and thus reduce the workload for the teacher, who can then devote more time to students who need special support. In this work, we are interested in intelligent tutors to help high-school students to solve questions of plane geometry. Unfortunately, in today's smart tutors, and especially in tutors for users who speak French, automatic knowledge extraction is left to the authors of the problem. They have to extract information from the problem and interpret it before entering it manually in the user interface, which is in the form of a predefined input structure in the tutor (for example, a list of hypotheses and conclusions). This may cause erroneous results due to a bad or incomplete interpretation. From this observation was born the motivation for our research, which is the creation of an automatic knowledge extractor from plane geometry problems written in French. This extractor will automate the addition of new problems in the tutor. This extractor is part of the QED-Tutrix project, which aims to create an intelligent tutor for plane geometry as taught in the Quebec school context at the secondary level.

The extractor we propose is based on the syntactic and grammatical analysis of the statement, which is combined with a set of rules to extract knowledge. The approach adopted consists in adapting to the French language an already existing extractor which has been designed to deal with multiple choice questions written in English. This adaptation was evaluated on the questions translated into French and then on a set of questions drawn from Quebec textbooks. Our approach, which is based on a better formal representation of the output of the extractor (i.e. encoding the relationships extracted from the question), gives good results, not only for multiple choice questions, but also for questions inspired by Quebec textbooks.

TABLE DES MATIERES

REMERCIEMENTS.....	iii
RÉSUMÉ	iv
ABSTRACT.....	v
TABLE DES MATIERES	vi
LISTE DES TABLEAUX.....	ix
LISTE DES FIGURES	xi
LISTE DES SIGLES ET ABRÉVIATIONS	xii
LISTE DES ANNEXES	xiii
CHAPITRE 1 INTRODUCTION.....	1
1.1. Contenu du mémoire	5
CHAPITRE 2 REVUE DE LITTÉRATURE	7
2.1. Concepts et outils relatifs à l'extraction des connaissances	7
2.2. Extraction de connaissances à l'aide d'un modèle de règles : LIM-G [1]	9
2.3. Extraction des connaissances basée sur des techniques de traitement de la langue naturelle.....	13
2.3.1. Compréhension et formalisation automatiques des problèmes de géométrie à l'aide du modèle syntaxique-sémantique S ² [5]	13
2.3.2. Résolution de questions de géométrie SAT en combinant la compréhension du texte et l'interprétation des figures : GEOS	18
CHAPITRE 3 MÉTHODOLOGIE.....	35
3.1. Création du corpus et choix des énoncés de géométrie.....	35
3.1.1. Corpus initial.....	35
3.1.2. Corpus enrichi.....	35
3.2. Adaptation de GEOS pour l'extraction des connaissances des énoncés québécois..	38
3.3. Validation de l'extraction.....	39
3.3.1 La performance de l'extraction (rappel et précision)	39

3.3.2.	La résolution des problèmes	40
3.3.3.	Validation finale.....	41
CHAPITRE 4 ADAPTATION DE LA MÉTHODE GEOS POUR DES PROBLÈMES ÉNONCÉS EN FRANÇAIS		43
4.1.	GEOSFR: adaptation de GEOS aux questions SAT traduites en langue française ..	43
4.1.1.	Traduction et annotation des questions SAT	43
4.1.2.	Adaptation du modèle d'identification des concepts et des relations dans GEOS pour le français (GEOSFR)	45
4.1.3.	Résultats de GEOSFR.....	46
4.2.	L'extraction des hypothèses et des conclusions des questions tirées des manuels scolaires québécois ou choisies par les didacticiens	48
4.2.1.	Résultats d'extraction de GEOSFR	48
4.2.2.	Insuffisance du formalisme Ω pour représenter les hypothèses et les conclusions	50
CHAPITRE 5 CRÉATION DE L'EXTRACTEUR GEOSFR++		57
5.1.	Les schémas de la hiérarchie des classes (taxonomie).....	57
5.2.	La liste des prédicats et des fonctions de $\Omega +$	60
5.3.	Description de l'algorithme GEOSFR++ de l'extraction basée sur la $\Omega +$ avec des exemples.....	60
5.3.1.	Génération des étiquettes	61
5.3.2.	Génération des prédicats et des fonctions unaires	63
5.3.3.	Génération des prédicats et des fonctions binaires	65
5.3.4.	Génération des prédicats et des fonctions ternaires	67
5.3.5.	Extraction des Conclusions.....	68
5.3.6.	Extraction des hypothèses implicites et optimisation de l'ensemble des hypothèses extraites	70
5.3.7.	Adaptation de l'extracteur au générateur de preuve QED-Tutrix.....	81
CHAPITRE 6 ÉVALUATION		86

6.1. La validation croisée (rappel et précision)	86
6.2. Résultats sur l'ensemble de test	89
6.2.1. Rappel et précision.....	89
6.2.2. La résolution par QED-Tutrix.....	90
CHAPITRE 7 DISCUSSION	92
CHAPITRE 8 CONCLUSION	95
BIBLIOGRAPHIE	96
ANNEXE	101

LISTE DES TABLEAUX

Tableau 1 : un exemple d'une démonstration possible du problème ci-haut.....	3
Tableau 2 : Les trois représentations du problème : « Mr. Wang bought a rectangular orchard in a small town in central Philadelphia. He measured this land with a ruler and got 1 km for the length and 500 m for the width. Please find the area of this orchard in square meters. » ...	9
Tableau 3 : un exemple de quelques catégories de problème liées aux concepts de carré et de rectangle [1].	10
Tableau 4 : les questions pour diagnostiquer la compréhension de l'étudiant générées pour le problème ci-haut [1].....	13
Tableau 5: Un exemple de problème de géométrie et le résultat de l'analyse et de l'annotation. (a) Un exemple de problème de géométrie ; (b) le résultat de la segmentation en mots et de l'étiquetage POS; (c) le résultat de la détection des limites de la phrase, chaque ligne est une phrase distincte [5].	15
Tableau 6 : Le résultat de l'identification des éléments et des relations géométriques présents dans l'exemple du tableau 4 [5].	16
Tableau 7 : L'extraction des relations finales de premier ordre présentes dans les phrases du tableau 5 [5].	16
Tableau 8 : les résultats de la résolution des questions.....	17
Tableau 9 : les résultats de la performance de l'extraction.....	17
Tableau 10 : Résultats de la résolution de GEOS des questions SAT officielles	31
Tableau 11 : Résultats de la performance de l'extraction.....	31
Tableau 12 : Comparaison entre le nombre de concepts géométriques présents dans le corpus initial et enrichi	36
Tableau 13 : Comparaison des résultats de GEOS sur les questions SAT en anglais avec ceux de GEOSFR sur les mêmes questions traduites en français	47
Tableau 14 : Résultats de GEOSFR en fonction du modèle de la classification	47
Tableau 15 : Résultats de GEOSFR sur le corpus MSD.....	49
Tableau 16 : les concepts identifiés dans les canevas des exercices tirés des manuels scolaires	55
Tableau 17 : une démonstration du second exemple.	72
Tableau 18 : Une partie de l'ensemble des règles pour l'extraction des hypothèses implicites (ERHI).....	74

Tableau 19 : Une partie de l'ensemble des règles pour l'extraction des hypothèses implicites (ERHI) concernant les conjonctions de coordination	80
Tableau 20 : Statistique sur les concepts géométriques de QED-Tutrix et $\Omega +$	84
Tableau 21 : Les concepts géométriques présents dans $\Omega +$ mais pas dans QED-Tutrix.	84
Tableau 22 : comparaison des résultats de GEOSFR et GEOSFR++	86
Tableau 23 : Comparaison entre GEOSFR++ sans et avec le processus de la finalisation par conjonction de coordination amélioré	87
Tableau 24 : Comparaison des résultats de GEOSFR++ avant et après la correction des angles	88
Tableau 25 : Les résultats de GEOSFR++ sur l'ensemble de test	89
Tableau 26 : Résultats de la résolution par QED-Tutrix	90
Tableau 27 : Comparaison des résultats avant et après le traitement des hypothèses implicites	91

LISTE DES FIGURES

Figure 1 : la figure illustrative de l'exemple.....	2
Figure 2 : analyse syntaxique de SyntaxNet de la phrase "Alice aime Bob"	8
Figure 3 : La représentation des connaissances du catégorie 2.a à l'aide de InfoMap [1]	11
Figure 4 : un exemple d'un arbre concept-attribut d'un problème de catégorie 2.a. [1]	11
Figure 5 : Question (colonne à gauche) et l'interprétation par GEOS (colonne à droite) [6]..	20
Figure 6 : l'arbre de dépendance syntaxique de l'expression 'L'angle ABC est aigu'.	25
Figure 7 : la taxonomie du formalisme Ω	28
Figure 8 : des exemples des axiomes utilisés dans [18].....	32
Figure 9 : un exemple d'un graphe HPDIC	41
Figure 10 : annotation d'une question SAT originale en anglais avec le formalisme Ω de GEOS.	44
Figure 11 : annotation d'une question SAT traduite en français avec la logique de premier ordre de GEOS.	45
Figure 12 : exemple d'un canevas avec un exercice tiré du manuel Sommet	55
Figure 13 : vue d'ensemble de l'algorithme GEOSFR++	61
Figure 14 : L'analyse syntaxique de l'énoncé 'L'angle ABC est aigu et l'angle CBD est droit' obtenue par SyntaxNet.....	64
Figure 15 : Figure illustrative de l'exemple.....	71
Figure 16 : Un exemple de représentation figurale de deux droites parallèles qui intersectent une troisième droite.....	82
Figure 17 : Exemple de représentation figurale d'angles : $angle([C,E],A,[D,B])$ en vert, $angle([D,B],A,[F])$ en rouge et $angle([F],A,[G])$ en bleu.....	83
Figure 18 : Le nombre d'hypothèses par rapport au nombre de points	94

LISTE DES SIGLES ET ABRÉVIATIONS

GEOS : Un extracteur automatique des connaissances des questions de géométrie plane à choix multiple SAT en anglais.

GEOSFR : l'extracteur GEOS adapté aux questions SAT traduites en français.

GEOSFR++ : l'extracteur GEOS adapté aux questions de géométrie québécoises du niveau secondaire.

POS : Part of speech tagging ou Etiquetage grammatical. Consiste à associer aux mots d'un texte les fonctions grammaticales correspondantes, par exemple, nom, verbe, adjectifs, etc.

TALN : Traitement automatique de la langue naturelle. Il s'agit d'un domaine multidisciplinaire qui implique l'informatique et la linguistique. Il est utilisé dans plusieurs applications, par exemple, l'extraction de l'information, la traduction, la reconnaissance vocale, etc.

LISTE DES ANNEXES

Annexe 1 : la hiérarchie des concepts de la logique Ω + ...	101
Annexe 2 : la listes des prédicats et des fonctions de la logique Ω	102
Annexe 3 : les prédicats et les fonctions de la logique Ω + (les concepts en gras sont les arguments présents dans les prédicats et les fonctions)	104
Annexe 4 : La liste des questions pour la validation finale	108

CHAPITRE 1 INTRODUCTION

Les systèmes tutoriels intelligents sont des outils informatiques d'apprentissage qui permettent d'assister un apprenant dans une tâche donnée, en lui offrant des rétroactions adéquates sur ses actions. En géométrie plane, les systèmes tutoriels auxquels nous nous intéressons sont les systèmes qui offrent un soutien intelligent aux apprenants dans la résolution d'un problème de démonstration. Plusieurs tuteurs de l'état de l'art sont conçus pour effectuer cette tâche. Nous citons par exemple Geometry Tutor [37], ANGLE [38], Baghera [39], DEFI [40], Cabri-Euclide [41], Mentoniez [42], Géométrie [43] et Turing [44]. Notre extracteur s'inscrit dans le projet de développement du tuteur intelligent QED-Tutrix [10] [45]. La particularité de QED-Tutrix est qu'il offre la possibilité à l'apprenant de faire une exploration libre des problèmes. C'est-à-dire, ce tuteur ne se restreint pas à une solution particulière d'un problème donné, mais encourage à explorer un espace de solutions plus grand. Les problèmes traités par QED-Tutrix sont ceux enseignés dans le contexte scolaire québécois au niveau secondaire. Dans un premier temps, l'enseignant insère dans QED-Tutrix l'ensemble des inférences qui composent la démonstration d'un problème donné. Ces inférences sont combinées pour construire le graphe de toutes les solutions possibles. Chaque nœud d'inférence contient des antécédents (hypothèses), une justification (propriété ou définition) et un conséquent (conclusion). Les conséquents peuvent être utilisés comme des hypothèses d'autres nœuds d'inférence, ce qui permet de compléter le graphe des solutions. Ensuite, grâce à ce graphe, le tuteur est capable de suivre la progression de l'élève dans sa résolution du problème afin de proposer des aides pertinentes en cas de blocage. L'insertion manuelle par l'auteur du problème de l'ensemble des inférences utilisées dans les démonstrations d'un problème donné représente une tâche laborieuse et difficile de mettre en œuvre à grande échelle. Cette opération a été partiellement automatisée dans [45], où on propose une approche pour générer automatiquement le graphe qui représente l'ensemble des solutions. Dans ce travail, nous nous intéressons à l'autre partie du problème, soit l'extraction automatique des hypothèses et conclusions qui servent à alimenter la génération de ce graphe.

Tout d'abord, il faut comprendre que la résolution des questions de géométrie plane est constituée de trois étapes principales :

- L'extraction des connaissances, qui consiste à extraire les hypothèses et les conclusions à partir de l'énoncé du problème.

- La définition du référentiel, qui consiste à déterminer l'ensemble des propriétés et des définitions de concepts qui seront utilisées dans la démonstration.
- La démonstration, qui consiste à appliquer une succession de règles qui permettent d'aller des hypothèses aux conclusions

Pour illustrer, prenons l'exemple suivant :

Soit un triangle ABC. AM est la médiane de ABC issue de A. Le point M est sur la droite BC. Les mesures des segments AM, BM et MC sont égales. La mesure de l'angle ACM est égale à 24. Trouver la mesure de l'angle ABC.

La figure 1 présente une figure illustrative de cet exemple.

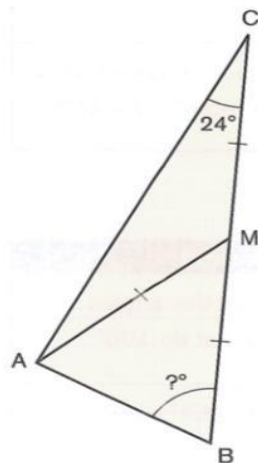


Figure 1 : la figure illustrative de l'exemple.

L'étape d'extraction des connaissances résulte en l'ensemble d'hypothèses suivant:

- ABC est un triangle
- AM est la médiane du triangle ABC issue de A
- $AM = BM = MC$
- M appartient à BC
- $\text{Mesure}(\text{ACM}) = 24 \text{ degrés}$

La conclusion est :

- La mesure de l'angle ABC.

Finalement, le tableau 1 présente une démonstration possible pour cet exemple.

Tableau 1 : un exemple d'une démonstration possible du problème ci-haut.

Affirmation	Justification
Mesure(CAM) = 24	AMC est un triangle isocèle. Il y a au moins 2 angles isométriques dans un triangle isocèle. Ces deux angles sont CAM et ACM (propriété de la médiane). CAM a donc la même valeur que ACM, soit 24.
Mesure(AMC) = 132	$ACM + CAM + AMC = 180$ dans le triangle ACM (La somme des angles dans un triangle)
Mesure(AMB) = 48	BMC est un angle plat. Sa mesure est 180. Donc $AMB + AMC = 180$
Mesure(ABM) = Mesure(BAM) = 66	Le triangle AMB est isocèle. Les angles ABM et BAM sont isométriques. Donc $2 \cdot ABM + 48 = 180$

Le référentiel est les propriétés et les définitions qui apparaissent dans la colonne droite du tableau 1, dont on se sert comme justification.

Nous remarquons dans le tableau 1 que certaines des justifications utilisent des hypothèses qui ne sont pas présentes explicitement dans le texte mais qui sont nécessaires pour la démonstration : les triangles ACM et AMB ne sont pas présents dans le texte mais ils apparaissent dans la démonstration. Ce type d'hypothèse est ce que nous appelons une hypothèse *implicite*. L'extraction doit tenir compte de ce genre d'hypothèses qui n'apparaissent pas explicitement dans le texte mais nous pouvons les déduire. Par exemple, nous pouvons déduire le triangle ACM du fait que les points A, C et M sont trois points distincts et reliés par des droites AC, CM et AM.

La majorité des tuteurs ou systèmes de construction géométrique actuels comme GeoGebra [7], GeoProof [8], Geometry Expert[9] et QED-Tutrix [10] ne contiennent pas un module pour l'extraction automatique des connaissances. Cette tâche est laissée à l'auteur du problème, qui doit extraire les informations du texte et les interpréter afin de les saisir manuellement dans l'interface utilisateur, qui est sous forme d'une structure d'entrée prédéfinie (par exemple, une liste des hypothèses et des conclusions). En ce qui concerne le français, il n'existe pas actuellement un extracteur de connaissances pour la géométrie plane telle qu'elle est enseignée aux élèves de niveau secondaire au Québec. Cette dépendance sur une entrée manuelle des

données peut entraîner des résultats erronés à la suite d'une mauvaise ou une incomplète interprétation de l'utilisateur.

L'extraction des connaissances dans des domaines spécifiques comme la géométrie, les mathématiques en général, la physique, ou la chimie est une tâche plus compliquée que l'extraction utilisée dans les documents plus généraux, comme les textes des journaux. L'extraction des connaissances dans un domaine comme les mathématiques consiste à utiliser les connaissances spécifiques du domaine pour isoler et repérer les informations clés dans les données. Ce qui nécessite de recourir à des experts et des spécialistes du domaine pour repérer ces informations.

En géométrie plane [1,2,4,5,6], ces informations clés peuvent être décrites sous forme d'ontologies qui définissent l'ensemble des classes et des relations géométriques possibles que nous pouvons trouver dans un énoncé. Par exemple, en ce qui concerne les classes géométriques, nous pouvons définir que la classe *quadrilatère* est une sous-classe de *polygone*. Un autre exemple concernant les relations géométriques, la relation *EstBissectrice(DB, ABC)*, qui est une instance de la relation *EstBissectrice(droite, angle)*, précise que la droite *DB* est une bissectrice de l'angle *ABC*.

En général, dans l'état de l'art, l'extraction des connaissances en géométrie plane [1,2,4,5,6] est basée sur :

- Un ensemble de patrons : cela consiste à repérer dans le texte certains patrons pour trouver les relations géométriques. Prenons comme exemple le patron suivant : « les variables qui sont précédées par le mot 'cercle' sont de classe *cercle* ». Ce patron permettra de repérer dans l'expression suivante 'Soit un cercle *O*.' la relation *EstCercle(O)*.
- L'analyse syntaxique : cela consiste à faire une analyse syntaxique de l'énoncé afin d'extraire la sémantique. Il s'agit ici d'une application du traitement automatique de la langue naturelle.

Nous expliquons plus en détail ces deux techniques dans le chapitre suivant.

De ces constatations naît l'objectif principal de notre travail de recherche.

Objectif principal :

Créer un extracteur automatique capable de dégager l'ensemble des **hypothèses** et des **conclusions** d'un énoncé de problème de géométrie plane en français. Par la suite, celles-ci seront utilisées par un générateur de preuve pour la résolution du problème. Les problèmes traités sont ceux présents dans les manuels du niveau secondaire au Québec.

Les extracteurs de connaissances [4,5,6] qui utilisent l'analyse syntaxique sont plus flexibles par rapport aux différentes façons d'énoncer les problèmes de géométrie. Cependant, jusqu'à maintenant, cette technique n'est utilisée qu'avec l'anglais [6], le chinois [5] et le taïwanais [1]. De ce fait, nous avons décidé d'appliquer la même technique sur des énoncés en français. Ceci nous amène à notre principale hypothèse :

Hypothèse 1 : Il est possible, à partir d'un énoncé de problème de géométrie plane typique de ce qui est utilisé au Québec pour les élèves du secondaire, d'en extraire automatiquement les connaissances permettant de générer les différentes démonstrations possibles pour résoudre le problème. Il faut noter ici que nous supposons que toutes les données nécessaires pour la résolution se trouvent dans l'énoncé.

En plus de cette hypothèse, qui est la principale de notre travail, nous croyons que l'analyse syntaxique, qui a montré ses bénéfices dans plusieurs domaines, peut être une partie importante de la solution, ce qui nous amène à l'hypothèse suivante :

Hypothèse 2 : Une analyse syntaxique des énoncés de langue française rend l'extraction des connaissances plus flexible aux différentes façons d'énoncer un problème de géométrie plane.

1.1. Contenu du mémoire

Ce mémoire commence par une revue de littérature (chapitre 2), où nous présentons les différents travaux effectués jusqu'à présent dans l'extraction des connaissances de problèmes de géométrie plane. Le chapitre 3 présente la méthodologie que nous avons suivie pour créer notre extracteur. Le chapitre 4 présente l'extracteur GEOSFR, notre premier extracteur, qui est une adaptation de l'extracteur GEOS [6] pour les questions SAT traduites en français. Le

chapitre 5 présente notre deuxième extracteur GEOSFR++ pour les questions de géométrie québécoises. Ce deuxième extracteur vient de la nécessité de changer la représentation utilisée dans GEOS et GEOSFR pour l'extraction, afin de la rendre applicable aux questions du niveau secondaire québécois. Le chapitre 6 présente les résultats de l'évaluation de GEOSFR++. Le chapitre 7 présente la discussion. Enfin, le chapitre 8 présente la conclusion.

CHAPITRE 2 REVUE DE LITTÉRATURE

Ce chapitre présente les différents travaux de l'état de l'art qui ont servi de base à notre travail. Les travaux qui font l'extraction des connaissances des problèmes de géométrie plane se divisent en deux types : le premier type [1] est basé sur des règles et des patrons pour l'extraction. Le deuxième type [4,5,6] est basé sur l'analyse syntaxique.

Avant d'aborder ces approches d'extraction, nous présentons quelques concepts de traitement automatique de la langue naturelle qui sont cités à plusieurs reprises dans ce mémoire. Puis, nous présentons quelques outils externes déjà existants dont nous nous sommes servis pour faire l'analyse syntaxique.

2.1. Concepts et outils relatifs à l'extraction des connaissances

Les concepts présentés ici sont liés à l'**analyse syntaxique**. Notre extracteur est basé sur celle-ci afin de dégager les hypothèses et les conclusions des problèmes donnés.

Syntaxe [47] : En linguistique, il s'agit de la façon dont les mots sont combinés pour former une phrase dans une langue. Elle signifie aussi de savoir si une séquence de mots respecte les règles d'une grammaire d'une langue donnée.

Analyse syntaxique [48] : Consiste à expliciter les relations qui existent entre les mots sous forme d'un **arbre syntaxique**. Le résultat de cette analyse est utilisé comme base pour l'analyse sémantique et la correction grammaticale.

Arbre syntaxique ou arbre de dépendance [15] : Il s'agit d'une représentation des relations syntaxiques qui relient les mots d'une phrase sous forme d'un arbre (voir l'exemple de la figure 2).

POS ou part of speech [15] : Représente la nature ou la fonction grammaticale d'un mot dans une phrase (par exemple : nom, verbe, déterminant, etc.).

Coréférence : Lorsque deux expressions ou plus réfèrent à la même entité. Par exemple, “*Soit le triangle ABC rectangle en A. d est la bissectrice de **son angle droit***”, dans cette expression **son angle droit** désigne l’**angle BAC** du triangle ABC de la phrase précédente.

Les outils déjà existants sont utilisés afin d’effectuer l’analyse syntaxique des problèmes. Nous nous intéressons uniquement aux résultats de l’analyse de ces outils, c’est-à-dire, sans rentrer dans les détails du fonctionnement de ces derniers.

SyntaxNet [15] : Il s'agit d'un service de Google, qui est un réseau neuronal qui a été développé pour faire l’analyse syntaxique en plusieurs langues. Étant donnée une phrase à l’entrée de ce système, il donne à la sortie l’étiquetage **POS (part of speech)** de chaque mot de la phrase et les relations syntaxiques entre les mots sous forme d’un **arbre de dépendance syntaxique**. Prenant l’exemple de la figure 2 “Alice aime Bob” : l’analyse syntaxique donnera le graphe qui y est illustré. L’étiquette POS des mots *Alice* et *Bob* est *nom* et celle de *aime* est *verbe*. Le verbe *aime* est la racine de la phrase, *Alice* est le sujet (nsubj) de *aime* et *Bob* est le complément d’objet direct (dobj) de *aime*.

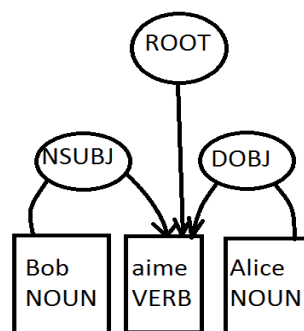


Figure 2 : analyse syntaxique de SyntaxNet de la phrase “Alice aime Bob”

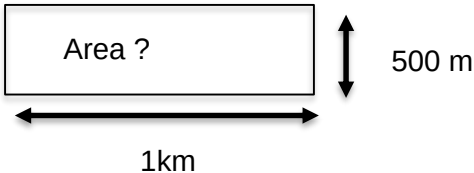
Stanford parser [14] : Est un autre analyseur syntaxique comme SyntaxNet utilisé dans GEOS [6]. Cependant, Stanford parser ne produit pas l’arbre de dépendance pour la langue française, ce qui nous oblige à le remplacer par SyntaxNet.

2.2. Extraction de connaissances à l'aide d'un modèle de règles : LIM-G [1]

Après que nous avons introduit, ci-haut, les concepts et les outils de traitement automatique de la langue naturelle qui sont abordés dans cette section et la section suivante, nous présentons les principaux travaux de l'état de l'art en relation avec l'extraction de connaissances des problèmes de géométrie plane.

Learner-initiating instruction model (LIM-G) est un extracteur de connaissances conçu pour aider les étudiants de l'école primaire de Taïwan à comprendre les problèmes géométriques. À cette fin, LIM-G transforme les énoncés textuels en des représentations télégraphiques et graphiques (voir le tableau 2) qui sont plus faciles à comprendre par les étudiants. En fait, les expérimentations dans les recherches [49, 50, 51, 52] ont montré que l'utilisation de différentes formes de représentation aide l'étudiant à mieux comprendre les problèmes et améliore les performances de la résolution.

Tableau 2 : un exemple de trois représentations de LIM-G.

Représentation textuelle	Mr. Wang bought a rectangular orchard in a small town in central Philadelphia. He measured this land with a ruler and got 1 km for the length and 500 m for the width. Please find the area of this orchard in square meters.
Représentation graphique	Rectangular orchard 
Représentation télégraphique	Rectangular orchard ; length :1km ;width :500m ; find the area ?

L'algorithme pour produire les deux représentations graphiques et télégraphiques est constitué des cinq étapes suivantes :

1. La détermination de la catégorie du problème : LIM-G regroupe manuellement dans une liste l'ensemble des catégories de problèmes qui sont enseignés dans le niveau primaire à Taiwan. Voir le tableau 3.

Tableau 3 : un exemple de quelques catégories de problème liées aux concepts de carré et de rectangle [1].

Formes géométriques	Catégories
1. Carré	1.a. Find the area of a square given the length of its sides 1.b. Find the perimeter of a square given the length of its sides 1.c. Et ainsi de suite
2. Rectangle	2.a. Find the area of a rectangle given its length and width 2.c Et ainsi de suite

Dans la suite du texte, nous prenons comme exemple le problème du tableau 2.

2. La représentation des connaissances basée sur InfoMap [2] : InfoMap est une ontologie pour la représentation des données au sein de différents domaines. Elle permet la compression, la détection et l'extraction des motifs ou des structures dans un texte donné. Prenons par exemple la catégorie 2.a '*Find the area of a rectangle given its length and width*' du tableau 3, la représentation des connaissances pour cette catégorie est donnée dans la figure 3. Cette représentation est faite manuellement pour toutes les catégories de problème. Les connaissances sont représentées sous forme d'un graphe hiérarchique de nœud des *connaissances schématiques*, des *concepts problèmes* et des *connaissances linguistiques*. La *connaissance schématique* sera une des catégories du tableau 3. Chaque *connaissance schématique* contient un ensemble de *concepts problèmes* comme *Rectangle*, *Given length*, etc. Chaque *concept problème* contient à son tour une *connaissance linguistique* nécessaire pour détecter ce concept dans l'énoncé du problème. Les *connaissances linguistiques* sont sous forme des patrons sur les mots du texte. Prenons en exemple le *concept problème Given width*, les *connaissances linguistiques* contient le patron suivant : ([*lexique pour la largeur d'un rectangle*] : [*lexique pour les prépositions*] : [*lexique pour les nombres*] : [*lexique pour les unités de la largeur*])). Dans cet exemple, le *lexique pour la largeur d'un rectangle* est l'ensemble de mots qu'on peut trouver dans l'énoncé du problème qui représente la largeur comme le mot '*width*'. Le *lexique pour les unités de la largeur* contient des mots comme '*m*', '*km*', etc. Il en est de même pour les autres lexiques.

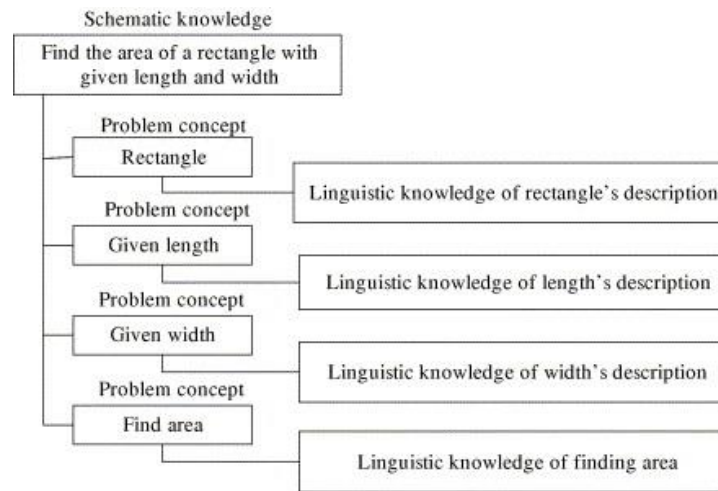


Figure 3 : La représentation des connaissances du catégorie 2.a à l'aide de InfoMap [1]

3. L'extraction des informations pour la génération d'un graphe concept-attribut : Grâce à l'étape précédente, le problème du tableau 2 est transformé automatiquement en arbre de concepts tel qu'illustré dans la figure 4. Pour cela, le système prend en entrée l'énoncé. Ensuite, les mots de l'énoncé déclenchent les patrons des *connaissances linguistiques* de chaque *concept problème*. Puis, le système cherche la *connaissance schématique* qui a tous les *concepts problème* déclenchés. Ainsi, le système détermine automatiquement la catégorie du problème.

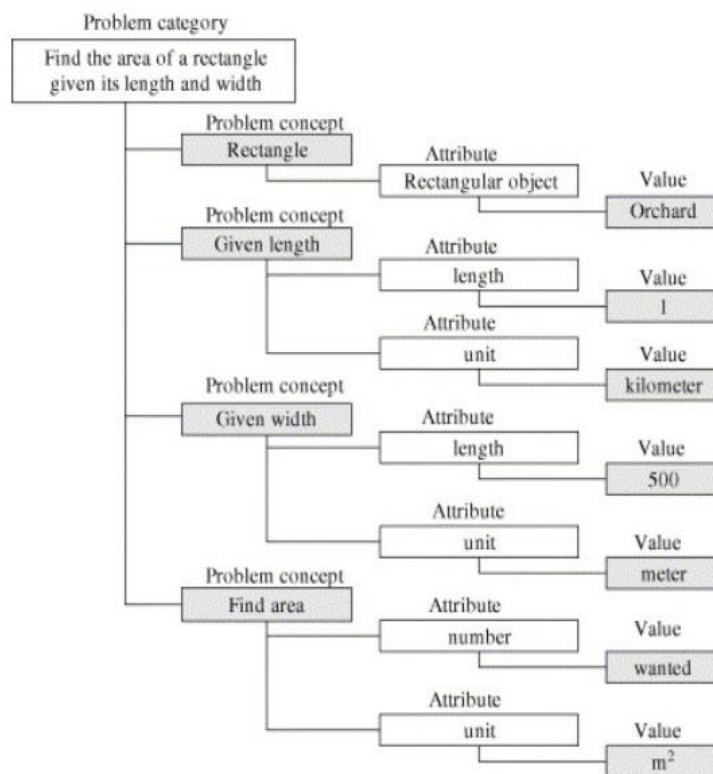


Figure 4 : un exemple d'un arbre concept-attribut d'un problème de catégorie 2.a. [1]

4. La génération des représentations télégraphique et graphique : LIM-G produit les deux représentations télégraphique et graphique simplifiées à partir de l'arbre concept-attribut.
5. L'évaluation de la compréhension de l'étudiant du problème : À partir des représentations télégraphique et graphique, LIM-G pose des questions à l'étudiant à propos des concepts et des attributs présents dans l'énoncé pour évaluer/diagnostiquer sa compréhension. Voir le tableau 4.

Pour résumer, LIM-G a pour but d'aider l'élève à comprendre la question en utilisant des représentations supplémentaires sous forme des représentations télégraphique et graphique. Ensuite, la compréhension de l'élève est évaluée à partir d'un ensemble de questions. L'extraction des connaissances effectuée par LIM-G pour créer les représentations supplémentaires pourrait être une piste prometteuse pour répondre aux objectifs de notre recherche. Cependant, cette méthode a plusieurs inconvénients :

- La méthode d'extraction des informations par les patrons telle que décrite ci-haut n'est efficace que sur les problèmes qui ne sont pas compliqués. On ne peut pas la généraliser à des problèmes qui contiennent plusieurs formes géométriques à la fois. Par exemple, dans un problème qui contient deux rectangles, on ne saura pas quelle longueur appartient à quel rectangle, puisqu'on n'utilise pas l'analyse syntaxique de l'énoncé. Il en va de même pour un problème qui contient par exemple un cercle et un rectangle, puisqu'on peut déclencher uniquement une catégorie avec le processus décrit ci-haut.
- L'évaluation de la performance dans cette recherche est effectuée en comparant les concepts et les attributs générés automatiquement avec ceux extraits manuellement pour chaque problème. Les connaissances extraites n'ont pas été évaluées par un générateur automatique de preuve. Autrement dit, nous ignorons si les données extraites sont suffisantes pour trouver automatiquement les démonstrations des questions.
- Les questions visées par LIM-G sont celles enseignées dans l'école primaire chinoise. Tous les manuels et cahiers d'exercices utilisés dans la collecte des questions sont conformes aux objectifs du cours spécifiés par le ministère de l'éducation chinois en 1996. Cependant, les questions visées par notre recherche sont celles enseignées dans l'école secondaire québécoise. Or ces deux types de questions sont différentes en ce qui concerne la langue et aussi les niveaux scolaires, et par conséquent les définitions et les

propriétés enseignées dans les deux contextes scolaires. De plus, la représentation des connaissances décrite ci-haut n'est pas compatible avec le niveau secondaire qui contient plusieurs formes géométriques à la fois.

Tableau 4 : les questions pour diagnostiquer la compréhension de l'étudiant générées pour le problème ci-haut [1].

Problem concept	Attribute and value/formula	Diagnosis question
Rectangle	Rectangular object: orchard	1. What is the shape of the orchard?
Given length of rectangle	The length of rectangle: 1 km	2. What is the length of the orchard?
	Unit of the length of rectangle: kilometer	3. What is the unit of the length of the orchard?

2.3. Extraction des connaissances basée sur des techniques de traitement de la langue naturelle

2.3.1. Compréhension et formalisation automatiques des problèmes de géométrie à l'aide du modèle syntaxique-sémantique S^2 [5]

Le modèle *syntaxique-sémantique* S^2 a pour but de comprendre les problèmes de géométrie plane présents dans les manuels scolaires chinois destinés aux élèves du secondaire, plus précisément, *Junior High School 8 et 9 grade : 12 à 14 ans*. Les problèmes utilisés dans les expérimentations sont tirés de l'édition PEP des manuels scolaires chinois et d'un livre de géométrie plane [11] et quelques problèmes dans [12].

L'extracteur utilise conjointement l'analyse syntaxique et un ensemble de patrons pour dégager les relations géométriques qui existent dans l'énoncé. Ces relations sont représentées à l'aide d'un ensemble de prédicats du premier ordre. L'extraction se fait en trois étapes :

- Analyse et annotation : Puisque le chinois est écrit sans séparation des mots, le texte est segmenté en mots et étiqueté par l'outil NLP ICTCLAS [13] conçu pour la langue chinoise. Voir l'exemple du tableau 5 (b). Une fois le texte segmenté en mot et étiqueté, il est segmenté en plusieurs phrases. Voir le tableau 5 (c). Cela se justifie par le fait que,

en général, chaque phrase du texte contient explicitement une seule relation géométrique à extraire.

- Identifications des éléments et des relations géométriques : Consiste à identifier les éléments et les relations géométriques présents dans le texte, en utilisant un dictionnaire de mots géométriques et les étiquettes ICTCLAS [13] :
 - Ce dictionnaire contient plusieurs éléments géométriques (comme *le parallélogramme*, *le cercle*, *le triangle*, etc.) et 48 relations géométriques (comme *eqDistance*, *midpoint*, etc.), voir le tableau 6.
 - ICTCLAS est outil de traitement automatique de la langue chinoise. Il permet de segmenter les phrases en plusieurs mots et aussi de trouver les étiquettes POS de ces derniers. D'une part, les variables géométriques (comme, ABCD, AD, etc.) dans l'étiquetage ICTCLAS sont étiquetées /x; d'autre part, les relations géométriques (comme equals, meet, etc.) sont étiquetées /v ou /n . Voir le tableau 5.

Ensuite, on calcule le nombre de lettre dans les variables géométriques étiquetées /x pour savoir sa nature. Par exemple, le nombre de lettre dans 'ABCD' est quatre donc il est quadrilatère.

Dans le tableau 6, on peut voir le résultat de cette identification des éléments et des relations géométriques de l'exemple du tableau 5.

- Extraction des relations : Un dictionnaire de patrons construit manuellement, appelé S^2 , est utilisé pour combiner les éléments géométriques avec les relations géométriques. Dans le tableau 7, les éléments et les relations identifiées précédemment sont combinés sous forme de relations de premier ordre.

Tableau 5: Un exemple de problème de géométrie et le résultat de l'analyse et de l'annotation.

(a) Un exemple de problème de géométrie ; (b) le résultat de la segmentation en mots et de l'étiquetage POS; (c) le résultat de la détection des limites de la phrase [5].

a	In quadrilateral ABCD, AD is equal to BC, and M, N are the midpoint of AB and CD respectively, AD and BC produced meet MN at point E and F. Prove angle DEN equals to angle CFN.
b	In/p quadrilateral/n ABCD/x ,/w AD/x equals/v to/p BC/x ,/w and/c M/x ,/w N/x are/v the/r midpoint/n of/p AB/x and/c CD/x respectively/d ,/w AD/x and/c BC/x produced/v meet/v MN/x at/p point/n E/x and/c F/x ./w prove/v angle/n DEN/x equals/v to/p angle/n CFN/x ./w
c	1- In/p quadrilateral/n ABCD/x ,/w 2- AD/x equals/v to/p BC/x ,/w 3- and/c M/x ,/w N/x are/v the/r midpoint/n of/p AB/x and/c CD/x respectively/d ,/w 4- AD/x and/c BC/x produced/v meet/v MN/x at/p point/n E/x and/c F/x ./w 5- (prove/v :/w) 6- angle/n DEN/x equals/v to/p angle/n CFN/x ./w

Tableau 6 : Le résultat de l'identification des éléments et des relations géométriques présents dans l'exemple du tableau 5 [5].

Phrases (voir les numéros dans la ligne c du tableau 5)	Eléments géométriques	Relation géométriques
1	(ABCD, quadrilatère)	Quadrilatère
2	(AD, line), (BC, line)	eqDistance
3	(M, point), (N, point), (AB, line), (CD, line)	midpoint
4	(AD, line), (BC, line), (MN, line), (E, point), (F, point)	intersect
6	(DEN, angle), (CFN, angle)	eqAngle

Tableau 7 : L'extraction des relations finales de premier ordre présentes dans les phrases du tableau 6 [5].

Phrases (voir les numéros dans la ligne c du tableau 5)	
1	Quadrilatère(ABCD)
2	eqDistance(AD,BC)
3	midpoint(M,AB) midpoint(N,CD)
4	intersect(AD,MN,E) intersect(BC,MN,F)
6	eqAngle(DEN, CFN)

Une fois les relations extraites, elles sont introduites dans le résolveur automatique JGEX [9]. JGEX est un système qui combine la génération automatique de preuve et la construction de la figure dynamique correspondante. Les méthodes utilisées dans le générateur de preuve sont la méthode de la base de données déductive [54], la méthode de Wu, la méthode de l'aire [56], la méthode de base de Grôbner et la méthode de plein angle [55].

L'extracteur est évalué selon deux aspects : la résolution du problème et la performance de l'extraction :

- La résolution du problème : Les relations extraites sont saisies dans le résolveur des problèmes de géométrie plane JGEX afin de calculer le nombre de problèmes résolus par rapport l'ensemble des problèmes de l'expérimentation. Dans le tableau 8, on présente les résultats de la résolution.

Tableau 8 : les résultats de la résolution des questions

Résultats	Résolues	Non résolues
Nombres de questions	119	11
Pourcentage	91.5%	8.5%

- La performance de l'extraction : Consiste à calculer la précision, le rappel et la F-mesure en comparant les relations extraites automatiquement avec les relations extraites manuellement.

Tableau 9 : les résultats de la performance de l'extraction

Résultats	Précision	Rappel	F-mesure
Relation unaire	1.0	0.990	0.995
Relation binaire	0.997	0.995	0.996
Relation Ternaire	0.979	0.963	0.971
Total	0.994	0.986	0.990

L'extracteur proposé ici est un candidat possible pour répondre à nos objectifs de départ. Contrairement à l'extracteur LIM-G qui se focalise plus sur la simplification et la représentation graphique du problème, l'extracteur S^2 est conçu pour fournir au résolveur les données nécessaires pour la résolution du problème. Cependant, cet extracteur a plusieurs inconvénients :

- Pour extraire les relations géométriques de l'énoncé, cette méthode segmente ce dernier en plusieurs phrases en utilisant les signes de ponctuation. Ensuite, chaque phrase est traitée séparément. On suppose ici que dans les problèmes traitables par cette méthode, chaque phrase contient une relation géométrique qui peut être extraite indépendamment

du reste des phrases. Cette technique a des limites dans le cas où les phrases ne sont pas indépendantes, c'est-à-dire, si elles contiennent des coréférences. Prenons en exemple l'énoncé suivant 'Soit un triangle ABC isocèle. Soit L la bissectrice issue du sommet A.' : dans ce cas, nous ne pouvons déduire que L est la bissectrice de l'angle ABC si on ne traite pas les deux phrases conjointement. Ce qui implique que les problèmes de ce genre, qui sont très présents dans les manuels scolaires québécois, ne seront pas traitables par cette méthode.

- De plus, si une phrase contient plus qu'une relation géométrique, cette méthode ne sera pas en mesure de les détecter toutes. Parce que, d'après le point ci-haut, chaque phrase doit contenir une seule relation. Prenons en exemple la phrase suivante 'La parallèle à la droite BC passant par I coupe la droite AC en J' : cette phrase contient deux relations géométriques $\text{Parallèle}(IJ,BC)$ et $\text{IntersectEn}(IJ,AC,J)$. Par conséquent, les problèmes de ce genre ne seront pas traitables par cette méthode.
- Cette méthode est incapable de traiter les problèmes géométriques de ce type '*Démontrer qu'un quadrilatère ayant 3 angles droits est un rectangle.*' où les éléments géométriques ne sont pas nommés.
- Cette méthode se restreint aux problèmes de démonstration où on a un ensemble d'hypothèses données et on cherche à prouver une conclusion connue dans le texte. Autrement dit, les problèmes de ce type '*ABCD est un losange de côté 60 cm. Trouver la mesure de la longueur BD*' ne sont pas traitables, parce que la conclusion, qui est ici la longueur de BD, est inconnue.

2.3.2. Résolution de questions de géométrie SAT en combinant la compréhension du texte et l'interprétation des figures : GEOS

Dans cette section, nous présentons en détail l'approche utilisée dans GEOS [6], qui constitue le point de départ de notre propre approche. En fait, comme nous le verrons plus tard, notre approche réutilise les principaux principes de l'approche, tout en y apportant des améliorations.

2.3.2.1. Description générale de GEOS

GEOS est un système de résolution des questions à choix multiple SAT de géométrie plane qui combine la compréhension du texte et l'interprétation de la figure illustrative. Le traitement des questions SAT par GEOS se fait en deux **étapes** : la première étape consiste à comprendre

l'énoncé en le transformant en des relations géométriques ; la deuxième étape consiste à résoudre les questions SAT en utilisant ces relations géométriques, qui résument les hypothèses et les conclusions qui se trouvent dans le texte et la figure. Dans la suite, nous nous intéressons ici uniquement à l'extraction des informations du texte sans tenir compte de l'extraction de la figure.

Les relations géométriques sont exprimées à l'aide d'une représentation formelle de premier ordre appelée Ω , qui contient :

- **Des constantes** : Il s'agit des nombres connus ou des entités géométriques avec des coordonnées connues.
- **Des variables** : Il s'agit des nombres ou des entités géométriques inconnus (p. ex. dans l'expression "*le triangle ABC*", ABC est une variable)
- **Des prédicats** : il s'agit des relations géométriques ou arithmétiques entre les **variables** et les **constantes** (p. ex. la relation $Perpendiculaire(AB,CD)$ qui indique que AB est perpendiculaire à CD)
- **Des fonctions** : il s'agit des fonctions de mesure liées aux **variables** et aux **constantes** (p. ex. la fonction $LongueurDe([AB])$ qui retourne la longueur du segment $[AB]$) ou des opérations arithmétiques (p. ex. la fonction $Somme(Mesure(ABC),Mesure(CBD))$ qui retourne la somme des deux angles ABC et CBD).

À l'aide de cette représentation, des questions comme celles de la colonne de gauche de la figure 5 peuvent être interprétées en des relations comme celles de la colonne de droite de la même figure.

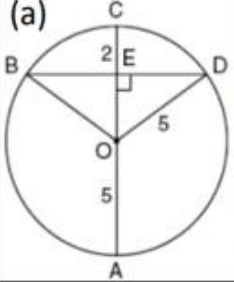
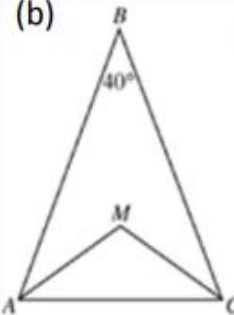
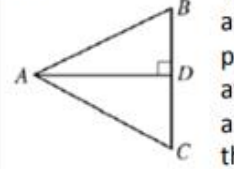
Questions	Interpretations
(a)  In the diagram at the left, circle O has a radius of 5, and $CE = 2$. Diameter AC is perpendicular to chord BD. What is the length of BD?	<i>Equals(RadiusOf(O), 5)</i> <i>IsCircle(O)</i> <i>Equals(LengthOf(CE), 2)</i> <i>IsDiameter(AC)</i> <i>IsChord(BD)</i> <i>Perpendicular(AC), (BD)</i> <i>Equals(what, Length(BD))</i> a) 12 b) 10 c) 8 d) 6 e) 4
(b)  In isosceles triangle ABC at the left, lines AM and CM are the angle bisectors of angles BAC and BCA. What is the measure of angle AMC?	<i>IsIsoscelesTriangle(ABC)</i> <i>BisectsAngle(AM, BAC)</i> <i>IsLine(AM)</i> <i>CC(AM, CM)</i> <i>CC(BAC, BCA)</i> <i>IsAngle(BAC)</i> <i>IsAngle(AMC)</i> <i>Equals(what, MeasureOf(AMC))</i> a) 110 b) 115 c) 120 d) 125 e) 130
(c)  In the figure at left, The bisector of angle BAC is perpendicular to BC at point D. If $AB = 6$ and $BD = 3$, what is the measure of angle BAC?	<i>IsAngle(BAC)</i> <i>BisectsAngle(line, BAC)</i> <i>Perpendicular(line, BC)</i> <i>Equals(LengthOf(AB), 6)</i> <i>Equals(LengthOf(BD), 3)</i> <i>IsAngle(BAC)</i> <i>Equals(what, MeasureOf(BAC))</i> a) 15 b) 30 c) 45 d) 60 e) 75

Figure 5 : Question (colonne à gauche) et l'interprétation par GEOS (colonne à droite) [6]

2.3.2.2. L'extraction des hypothèses et des conclusions

L'extraction des hypothèses et des conclusions se décompose en les étapes suivantes :

- Identification des concepts** : Elle consiste à repérer les mots désignant un concept géométrique et à les relier à plusieurs équivalents possibles dans la représentation formelle Ω (e.g. l'expression "l'angle ABC" sera étiqueté "l'angle(*IsAngle*) ABC(*angle/triangle*)", le mot 'angle' est identifié comme le prédicat *IsAngle*, et le mot 'ABC' est étiqueté comme une variable de type *angle* ou *triangle*). L'identification des concepts s'effectue par trois façons :
 - On compte le nombre de lettres majuscules des variables. Par exemple, le mot 'ABCD' qui contient quatre lettres majuscules sera étiqueté *quadrilatère*. Dans le cas du mot 'ABC', il est étiqueté *triangle* et *angle*. Ce qui explique la présence de deux étiquettes pour ce mot dans l'exemple ci-haut.

- En prenant en entrée des énoncés déjà étiquetés de l'ensemble d'entraînement, GEOS crée un dictionnaire qui relie les concepts de la représentation Ω avec les mots géométriques correspondants qu'on trouve dans le texte. Par l'exemple, si durant l'entraînement, GEOS a en entrée l'énoncé déjà étiqueté suivant '*le triangle(IsTriangle) ABC(triangle)*', à la sortie, les couples ('*triangle*', *IsTriangle*) et ('*ABC*', *triangle*) seront enregistrés dans ce dictionnaire. Dans le cas où on rencontre de nouveau les mots '*triangle*' et '*ABC*' dans un énoncé de l'ensemble de validation, ils seront automatiquement étiquetés, respectivement, *IsTriangle* et *triangle*. On explique dans l'étape prochaine plus en détail l'entraînement du modèle GEOS.
- Une troisième façon pour identifier les concepts est l'utilisation d'une liste de règles. Cela consiste à repérer dans le texte certains patrons pour trouver les concepts géométriques. Prenons l'exemple de la règle suivante : « les variables qui sont précédées par le mot 'carré' sont étiquetées *carré* ». Cette règle permettra d'étiqueter l'expression 'Soit un carré ABCD' par 'Soit un carré ABCD(*carré*)'.
- **Identification des relations** : Elle consiste à rassembler les prédicats et les fonctions avec leurs arguments qui sont les variables et les constantes. Une fonction peut être utilisée comme argument d'un prédicat ou d'une autre fonction, comme le montre la colonne *interprétation* de la figure 5. Un modèle de classification supervisé est utilisé. Il permet de décider si une relation rassemblée par le processus ci-haut est fausse ou vraie. Le vecteur caractéristique pour la classification est basé sur la distance dans le texte entre les mots, l'étiquetage POS (*part of speech*) et l'arbre de dépendance généré par Stanford Parser [14]. Plus précisément, l'identification des relations se décompose en deux étapes :
 - Combinaison des prédicats avec les constantes et les variables : Cette étape consiste à générer automatiquement toutes les combinaisons possibles entre les concepts trouvés dans l'étape précédente sans tenir compte du texte. Autrement dit, on génère des prédicats sans tenir compte ni de la position des mots dans le texte ni de la syntaxe. La seule condition qui doit être respectée ici est la compatibilité des variables et des constantes avec les arguments des fonctions et des prédicats, ce qu'on explique en détail dans la section 2.3.2.3. Prenons en exemple l'expression 'Les mesures des angles ABC et BCD sont égales à 20 et

30 degrés' : les concepts identifiés par l'étape précédente sont $\{ \text{Equals}(\text{number}, \text{number}), 20, 30, \text{Mesure}, \text{ABC}, \text{BCD} \}$. Cette étape va générer toutes ces combinaisons possibles $\{ \text{Equal}(\text{Mesure}(\text{ABC}), \text{Mesure}(\text{BCD})), \text{Equal}(\text{Mesure}(\text{ABC}), 20), \text{Equal}(\text{Mesure}(\text{BCD}), 30), \text{Equal}(20, 30) \}$. Dans cet exemple la fonction *Mesure* retourne un *number*, ce qui a permis de l'utiliser comme un argument du prédicat *Equals*.

Prenons le deuxième exemple suivant : 'A tangent line is drawn to circle O with radius of 5 and diameter of 7'. Ici, on génère les combinaisons suivantes $\{ \text{Equals}(\text{RadiusOf}(\text{O}), 5), \text{Equals}(\text{RadiusOf}(\text{O}), 7), \text{Equals}(\text{DiameterOf}(\text{O}), 5), \text{Equals}(\text{DiameterOf}(\text{O}), 7), \text{Equals}(\text{RadiusOf}(\text{O}), \text{DiameterOf}(\text{O})) \}$.

- Détermination des combinaisons correctes : on remarque dans les deux exemples précédents qu'il y a des combinaisons qui sont fausses et d'autres qui sont vraies. Pour les distinguer, un modèle de classification supervisé de *Regression Logistique* est utilisé. Un vecteur caractéristique est construit pour chaque combinaison issue de l'étape précédente en encodage à chaud (one-hot encoding), ce que nous expliquons par la suite :
 - Tout d'abord, on énumère tous les prédicats possibles. Chaque prédicat a une unique représentation binaire qui n'a qu'un seul chiffre 1 et tous les autres, la valeur 0. Ensuite, ce vecteur creux est concaténé au vecteur caractéristique qui est initialisé à une liste vide. Dans GEOS, le nombre de prédicats est 44, alors, la longueur de ce vecteur en encodage à chaud est aussi 44. A cette étape, la longueur du vecteur caractéristique est également 44.
 - De la même façon, nous représentons également les types des arguments par un vecteur en encodage à chaud. Ce vecteur est concaténé à la fin du vecteur caractéristique résultant de l'étape précédente. La longueur du vecteur représentant tous les types des arguments est 34 puisque le nombre de types d'arguments possible est 34. La longueur du vecteur caractéristique, à cette étape, est égale à $44 + 34 = 78$.
 - Après la génération de l'arbre de dépendance syntaxique pour chaque énoncé, on calcule la distance la plus courte entre le prédicat et ses arguments dans cet arbre. C'est-à-dire, on calcule le nombre de relations qui séparent le prédicat et les arguments dans l'arbre. L'arbre de

dépendance est celui généré par Stanford Parser [14]. De la même façon, cette distance est transformée en représentation en encodage à chaud et concaténée à la fin du vecteur caractéristique résultant de l'étape précédente.

- On calcule la distance entre le prédicat et ses arguments dans le texte, c'est-à-dire, le nombre de mots entre les deux plus 1. Ensuite, cette distance est transformée en une représentation binaire creuse et ajoutée à la fin du vecteur caractéristique résultant de l'étape précédente.
- Les étiquettes POS (nature grammaticale) du prédicat et de ses arguments sont également représentées en encodage à chaud et ajoutées à la fin du vecteur caractéristique.
- Les relations syntaxiques entre les prédicats et ses arguments sont également représentées en encodage à chaud et ajoutées à la fin du vecteur caractéristique. En fait, on commence avec le cas où la distance syntaxique est un où on liste la seule relation syntaxique qui relie le prédicat et l'argument, puis on passe au cas où la distance syntaxique est deux, et ainsi de suite. GEOS se restreint à traiter uniquement la distance syntaxique 1 et 2, c'est-à-dire, si la distance dans l'arbre de dépendance est supérieure strictement à 2, l'encodage à chaud de ces relations syntaxiques ne sera pas ajouté dans le vecteur caractéristique.

Ce processus est répété dans l'entraînement et la validation :

- L'entraînement : Ce modèle est entraîné sur une partie du corpus où les questions sont annotées avec les hypothèses et les conclusions. Pour ces questions d'entraînement, nous générons toutes les combinaisons possibles des prédicats avec les constantes et les variables, avec le processus décrit ci-haut. On génère ensuite les vecteurs caractéristiques pour ces combinaisons, on les appelle X_i . Puis, si la combinaison est présente dans l'annotation¹ de la question d'entraînement alors elle est correcte, ce qui veut dire qu'on lui assigne automatiquement la valeur 1, sinon, on lui assigne 0. On appelle cette valeur y_i . Ensuite, la matrice de $[X_i, y_i]_{i=1..N}$, où N est le nombre de combinaisons trouvées dans tout

¹ Les combinaisons générées automatiquement par l'extracteur sont comparées avec les annotations manuelles faites pour les questions de l'ensemble d'entraînement.

l'ensemble d'entraînement, est utilisé pour entraîner le classifieur à distinguer les combinaisons fausses et vraies. GEOS utilise 142 questions pour l'apprentissage et 44 questions pour la validation.

- La validation : Le modèle génère pour les questions de validation toutes les combinaisons possibles des prédicats avec les constantes et les variables. Ensuite, il génère les vecteurs caractéristiques pour chaque combinaison. Puis, le classifieur déjà entraîné produit pour chaque vecteur un score entre 0 et 1. Plus le score est proche de 1, plus la probabilité que la combinaison issue de l'étape précédente soit vraie est élevée.

Pour la validation, prenons l'exemple suivant 'L'angle ABC est aigu' : Les combinaisons générées dans l'étape précédente sont $\{IsAcute(ABC)\}$. La figure 6 présente l'arbre de dépendance syntaxique généré pour cet exemple.

- Le prédicat est *IsAcute* qu'on représente par l'encodage à chaud $[0, \dots, 0, 1_i, 0, \dots, 0]$ où 1 est à la i -ème position pour identifier le prédicat *IsAcute*. Le vecteur caractéristique est, alors, $VC = [0, \dots, 0, 1_i, 0, \dots, 0]$.
- Le type d'argument de *ABC* est *angle* qu'on représente par l'encodage à chaud $[0, \dots, 0, 1_j, 0, \dots, 0]$ où 1 est à la j -ème position qui correspond à *angle*. Ainsi, le vecteur caractéristique est $VC = VC + [0, \dots, 0, 1_i, 0, \dots, 0] = [0, \dots, 0, 1_i, 0, \dots, 0] + [0, \dots, 0, 1_j, 0, \dots, 0] = [0, \dots, 0, 1_i, 0, \dots, 0, 0, \dots, 0, 1_j, 0, \dots, 0]$.
- La distance dans l'arbre syntaxique entre le prédicat *IsAcute* et *ABC* est 2, car il y a deux relations qui les séparent NSUB et ACOMP. Le vecteur creux pour cette distance est $[0, 1, 0, \dots, 0]$ ($[1, 0, \dots, 0]$ représente la distance 1). Le vecteur caractéristique résultant est $VC = VC + [0, 1, 0, \dots, 0] = [0, \dots, 0, 1_i, 0, \dots, 0, 0, \dots, 0, 1_j, 0, \dots, 0, 0, 1, 0, \dots, 0]$.
- La distance dans le texte est 2, car il y a un mot entre *ABC* et *aigu* et on ajoute 1. Ainsi, le vecteur caractéristique est $VC = VC + [0, 1, 0, \dots, 0] = [0, \dots, 0, 1_i, 0, \dots, 0, 0, \dots, 0, 1_j, 0, \dots, 0, 0, 1, 0, \dots, 0, 0, 1, 0, \dots, 0]$.
- L'étiquette POS de *IsAcute* est ADJ, ce qui correspond à $[0, \dots, 0, 1_k, 0, \dots, 0]$, et celle de *ABC* est NOUN qui correspond à $[0, \dots, 0, 1_l, 0, \dots, 0]$. Le vecteur caractéristique résultant est $VC = VC + [0, \dots, 0, 1_k, 0, \dots, 0] + [0, \dots, 0, 1_l, 0, \dots, 0]$

- Les relations syntaxiques entre *IsAcute* et *ABC* sont NSUB $[0, \dots, 0, 1_m, 0, \dots, 0]$ et ACOMP $[0, \dots, 0, 1_n, 0, \dots, 0]$. Le vecteur caractéristique résultant est $VC = VC + [0, \dots, 0, 1_m, 0, \dots, 0] + [0, \dots, 0, 1_n, 0, \dots, 0]$.

Le modèle de classification supervisé déjà entraîné génère pour le vecteur VC le score 1.0, car la combinaison générée est vraie.

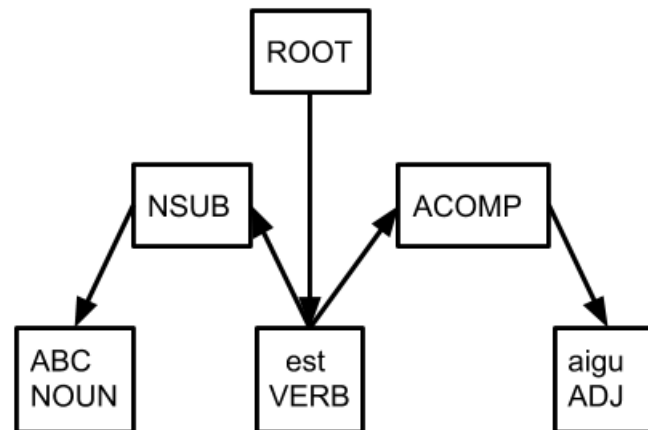


Figure 6 : l'arbre de dépendance syntaxique de l'expression 'L'angle ABC est aigu'.

- **Identification de l'ensemble final des prédicats :** Comme expliqué ci-haut, le classifieur déjà entraîné donne un score pour chaque combinaison d'un prédicat avec des constantes et des variables pour les questions de validation. On appelle ce score $S(c)$, où c est la combinaison d'un prédicat avec ses arguments. Cette étape consiste à choisir le sous-ensemble des combinaisons qui représente les hypothèses et les conclusions de la sortie. Pour un ensemble de combinaisons EC , le score $S(EC)$ est la somme des scores individuels de toutes les combinaisons : $S(EC) = \sum_{c \in EC} S(c)$. Ce score n'est pas normalisé. Pour choisir le sous-ensemble optimal de combinaisons, c'est-à-dire celui qui représente l'ensemble des hypothèses et des conclusions finales, GEOS utilise la fonction objective définie par l'équation suivante :

$$F(EC) = \lambda S(EC) + H(EC)$$

Où $H(EC)$ est une fonction de cohérence. Elle permet d'encourager l'extracteur à choisir un sous-ensemble de combinaisons qui couvrent le maximum de prédicats, de variables et de constantes présents dans l'énoncé et, en même temps, à éviter les redondances, c'est-à-dire, éviter de réutiliser les mêmes mots du texte pour identifier plusieurs concepts distincts. Prenons en exemple l'expression "Le **triangle rectangle**

ABC en A"., les relations qu'on peut identifier sont $\{IsTriangle[\mathbf{triangle}](ABC), IsRightTriangle[\mathbf{triangle} \ \mathbf{rectangle}](ABC)\}$. Les mots entre crochets indiquent les mots qui ont permis d'identifier les prédicats. On peut remarquer que le mot **triangle** a permis d'identifier deux prédicats $IsTriangle$ et $IsRightTriangle$. Le modèle de classification pénalise cette redondance par l'équation suivante :

$$H(EC) = N(EC) - R(EC)$$

Où $N(EC)$ est le nombre de prédicats et d'arguments couverts par le texte de l'énoncé. $R(EC)$ est le nombre de redondances dans EC . Prenons l'exemple ci-dessus avec l'ensemble de combinaisons trouvé $\{IsTriangle[\mathbf{triangle}](ABC), IsRightTriangle[\mathbf{triangle} \ \mathbf{rectangle}](ABC)\}$. $N(EC) = 3$ car on a deux prédicats $IsTriangle$ et $IsRightTriangle$, et une variable ABC dans l'ensemble de combinaisons. $R(EC) = 4$ car les mots **triangle** et ABC sont présents, tous les deux, deux fois dans l'ensemble des combinaisons. Pour revenir à $F(EC) = \lambda S(EC) + H(EC)$, λ est le paramètre de compromis entre S et H , qui est appris à travers l'ensemble de validation. L'algorithme *GREEDY_MAXIMIZATION* [4] (ci-dessous) décrit le processus pour choisir le sous-ensemble de prédicats optimal, en utilisant la fonction objective F [6] :

<i>GREEDY_MAXIMIZATION</i> : déterminer l'ensemble des prédicats optimal
Entrée : Ensemble des combinaisons EC, le score $S(c)$ pour chaque combinaison c Sortie : le sous-ensemble des combinaisons EC^* optimal Initialiser EC^* à l'ensemble vide Faire : $c^* = \underset{c \in EC \setminus EC^*}{argmax} F(EC^* \cup \{c\}) - F(EC^*)$; Ajouter c^* à EC^*; Répéter tant qu'il existe un $p \in EC \setminus EC^*$ tel que $F(EC^* \cup \{p\}) - F(EC^*) > 0$. Retourne EC^*

- **Identification des conclusions** : Les conclusions sont les prédicats identifiés précédemment dont un des arguments est inconnu. Prenons l'exemple suivant, pris de la figure 5 (a):

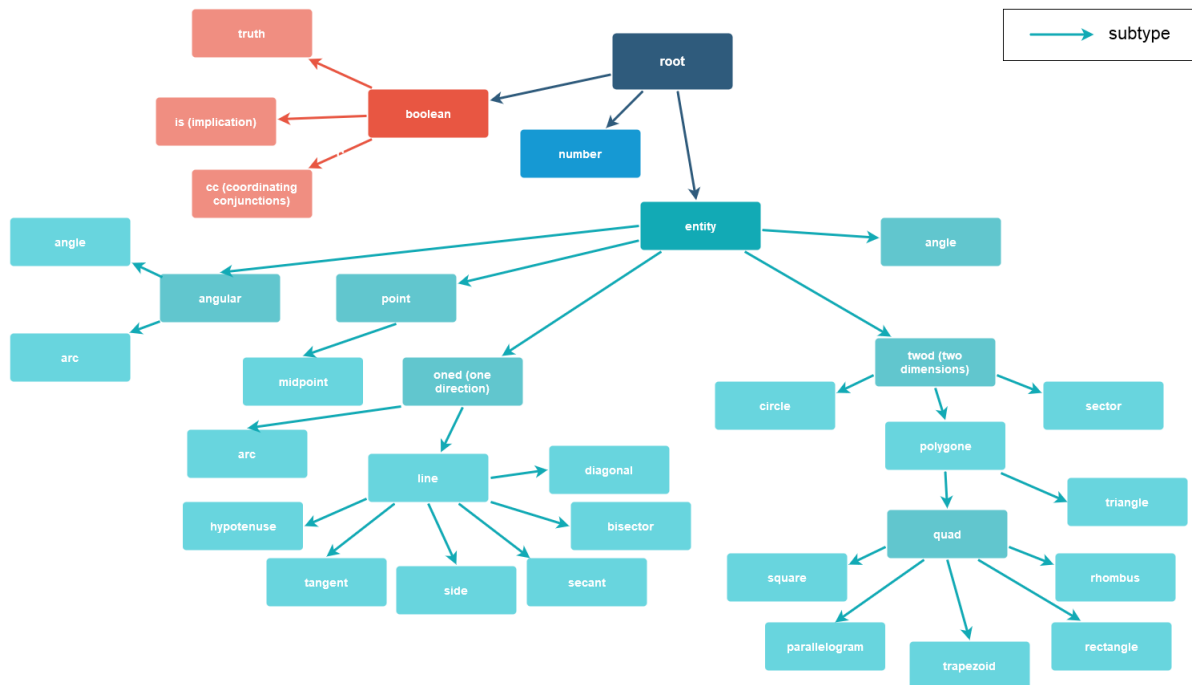
In the diagram at the left, circle O has a radius of 5 and $CE = 2$. Diameter AC is perpendicular to chord BD. What is the length of BD ?

On a la conclusion $Equals(what, Length(BD))$ où *what* est l'inconnue qui correspond à un nombre, qu'on cherche à trouver par le résolveur automatique. Nous décrivons en plus de détail le traitement des conclusions dans GEOS dans la section qui suit.

2.3.2.3. Description du formalisme de représentation Ω pour représenter les hypothèses et les conclusions

La taxonomie du formalisme de représentation Ω :

Dans la figure 7, on présente la taxonomie du formalisme Ω sous forme d'un arbre, c'est-à-dire, comment les concepts sont organisés dans GEOS. Les arêtes entre les concepts géométriques signifient qu'un concept est un sous-type d'un autre. Par exemple, le concept *square* est un sous-type du concept *quad*, ce dernier est un sous-type du concept *polygon*, etc. Cette taxonomie est utilisée dans la définition des prédicats et des fonctions : chaque prédicat ou fonction a des arguments dont la classe appartient à une des classes de la taxonomie de la figure 7. L'intérêt de cette taxonomie est qu'il permet de définir des prédicats qui sont plus généraux, ce qui permet d'éviter la redondance. Prenons en exemple le prédicat $IsInscribedIn(polygon, circle)$. Le premier argument est de classe *polygon*, ce qui signifie que n'importe quel concept de type *polygon* ou sous-classe de *polygon* (comme *quad*, *square*, etc.) peut être utilisé comme premier argument de $IsInscribedIn$. Prenons en exemple l'énoncé étiqueté suivant '*le quadrilatère(IsQuad) ABCD(quad) est inscrit(IsInscribedIn) dans le cercle(IsCircle) O(Point/Circle/Line)*', le prédicat $IsInscribedIn(polygon, circle)$ peut être combiné avec *ABCD*, puisque *quad* est sous-type de *polygon*. Par conséquent, en sortie, on a les prédicats $IsInscribedIn(ABCD, O)$ et $IsCircle(O)$. Il en va de même pour l'énoncé '*le triangle(IsTriangle) ABC(triangle) est inscrit(IsInscribedIn) dans le cercle(IsCircle) O(Point/Circle/Line)*' où on a en sortie les prédicats $IsInscribedIn(ABC, O)$ et $IsCircle(O)$, où $IsInscribedIn$ reste valable aussi pour le *triangle*, puisque ce dernier est sous classe de *polygon*. Dans les exemples ci-haut, on peut vérifier que *quad* et *triangle* sont des sous-classes de *polygon* en parcourant de bas en haut le graphe de la figure 7.

Figure 7 : la taxonomie du formalisme Ω

Les prédicats et les fonctions du formalisme Ω :

On différencie deux types de prédicats et de fonctions : binaires et unaires, selon qu'ils prennent en arguments un ou deux concepts. La différence entre les prédicats et les fonctions est que les prédicats sont de type booléen (les cases rouges dans la figure 7), alors que les fonctions sont de type non booléen. Les hypothèses et les conclusions sont représentées sous forme de prédicats. Ces prédicats prennent en argument les fonctions, les variables et les constantes. Le point commun entre ces derniers est qu'ils sont de type non booléen. Par exemple, l'expression "le périmètre du rectangle ABCD est égal à p" sera représentée sous la forme $Equal(PerimeterOf(ABCD),p)$. La fonction $PerimeterOf(ABCD)$ qui prend en argument un ABCD est de type *nombre*. Le prédicat $Equal(number,number)$ précise que $PerimeterOf(ABCD)$ est égal à p.

Dans le tableau de l'annexe 2, on regroupe l'ensemble des prédicats et des fonctions du formalisme Ω . La flèche des fonctions désigne la classe de la fonction. Par exemple, $AreaOf(twod) \rightarrow number$ signifie que la fonction $AreaOf$ est un *nombre*. La classe *twod* (*two dimensions*) regroupe toutes les formes géométriques de deux directions comme les cercles, les secteurs et les polygones. Dans ce cas les segments et les droites ne sont pas des sous-classes de *twod*.

Représentation des conclusions :

GEOS utilisent les fonctions *What()*->*number*, *Which()*->*number* et le prédicat *Find(number)*. Les deux fonctions *What* et *Which* sont combinées avec les prédicats de la liste ci-haut pour représenter les conclusions :

- **Which** : Pour extraire la conclusion d’une expression comme “Parmi les réponses suivantes, laquelle est la longueur de la ligne DB”, **Which** (en français, **laquelle**) représentera la longueur de la ligne DB qui est inconnue : *Equal(LengthOf(DB),Which)*.
- **What** : Pour extraire la conclusion d’une expression comme “Quelle est la mesure de l’angle PQS ?”, **What** (en français, **quelle**) représentera la mesure de l’angle PQS qui est inconnue : *Equal(MeasureOf(PQS),What)*.
- **Find** : Pour extraire la conclusion d’une expression comme “trouver la longueur de AB”, la conclusion sera formulée comme *Find(LengthOf(AB))*.

GEOS se restreint à extraire ces 3 types de conclusion que nous trouvons dans les questions SAT.

Finalisation des relations :

Certains types d’hypothèses et conclusions dans le texte ne sont pas extraites avec le processus décrit ci-haut, comme celles qui impliquent les **conjonctions de coordination** et les **implications**. Ce qui a amené GEOS à introduire des prédicats spéciaux *CC(root,root)* et *Is(root,root)* :

- **Conjonction de coordination** : Les expressions comme “AM et CM sont des bissectrices de BAC et BCA” ne peuvent pas être traitées directement par le processus d’extraction de GEOS décrit ci-haut (voir la section 2.3.2 : L’identification de l’ensemble final des prédicats). En fait, normalement, l’expression ci-dessus doit être représentée par les prédicats *BisectsAngle[bissectrices](AM,BAC)* et *BisectsAngle[bissectrices](CM,BAC)*. Cependant, un des deux prédicats n’est pas présent dans la sortie de l’extraction, parce que le *GREEDY_MAXIMAZATION* (voir la section 4.1) pénalise la redondance : le mot **bissectrices** est utilisé dans les deux prédicats identifiés ci-haut, ce qui n’est pas permis par *GREEDY_MAXIMAZATION*. Pour pallier ce problème, une représentation intermédiaire est utilisée *CC(root,root)*. Ainsi, l’expression “AM et CM sont des bissectrices de BAC et BCA” sera plutôt

représentée par les relations $BisectsAngle[bissectrices](AM,BAC)$, $CC(AM,CM)$ et $CC(BAC,BCA)$. Ce n'est pas l'unique résultat qu'on peut avoir, nous pouvons aussi avoir en sortie, $BisectsAngle[bissectrices](CM,BAC)$, soit $BisectsAngle[bissectrices](AM,BCA)$ ou encore $BisectsAngle[bissectrices](CM,BCA)$. Grâce aux relations intermédiaires CC , GEOS induit la relation $BisectsAngles(\{AM,CM\},\{BAC,BCA\})$. Un problème, qui n'est pas traité par GEOS, subsiste encore : cette relation n'implique pas directement $BisectsAngle[bissectrices](AM,BAC)$ et $BisectsAngle[bissectrices](CM,BAC)$, parce que dans $BisectsAngles(\{AM,CM\},\{BAC,BCA\})$ nous ne savons pas quelle droite est la bissectrice de quel angle. Autrement dit, l'ordre dans le texte des mots AM et CM (de même pour BAC et BCA) n'est pas utilisé dans GEOS pour savoir que AM est la bissectrice de BAC , et CM est la bissectrice de BCA . Ce problème est traité dans les améliorations que nous proposons dans la suite. Notons que le prédicat CC est un prédicat intermédiaire qui ne sera pas présent dans les résultats d'extraction finaux.

- **Implication** : Les rôles du verbe *être* dans les expressions 'C **est** un cercle', 'ABC **est** un triangle' et 'La mesure de AB **est** 3' ne sont pas les mêmes. On ne peut pas durant l'entraînement associer **est** aux trois prédicats différents $isCircle[est]$, $isTriangle[est]$, $Equal[est]$, etc, car les expériences ont montré que cela génère beaucoup d'erreurs dans l'identification des concepts.

Pour pallier ce problème, GEOS recourt au prédicat $Is(root,root)$. $root$ est la racine de la taxonomie du Ω . Ce qui veut dire que Is accepte tous les types d'arguments. On a deux cas d'utilisation de ce prédicat :

- Dans le cas où **est** signifie *égal*, comme dans l'exemple suivant 'La mesure de AB **est** 3', alors $Is(root,root)$ est transformé en $Equal(nombre, nombre)$. Par conséquent, $Is(Mesure(AB),3)$ est transformé en $Equal(Mesure(AB),3)$. Pour reconnaître ce cas, les deux $root$ du prédicat $Is(root,root)$ sont de type *number* : dans l'exemple précédent $Mesure(AB)$ et 3 sont de type *number*.
- Sinon, Is est transformé en des relations de type ($IsCircle$, $IsTriangle$ $IsSquare$, etc). Dans les exemples 'C **est** un cercle', 'ABC **est** un triangle', $Is(C,cercle)$ et $Is(ABC,triangle)$ sont transformés respectivement en $IsCircle(C)$ et $IsTriangle(ABC)$.

De même, Is est un prédicat intermédiaire qui ne sera pas présent dans les résultats d'extraction finaux.

2.3.2.4. L'évaluation de GEOS

Le processus de l'évaluation de l'extraction est semblable à celui utilisé par S^2 [5] décrit ci-dessus. En fait, il est évalué en termes de performance pour la résolution du problème, d'une part, et l'extraction d'autre part. Les problèmes utilisés dans l'expérimentation sont tirés des anciens examens SAT du College Board [16]. D'autres problèmes de géométrie plane sont pris de [17].

- Résultats de la résolution du problème : 55 questions extraites des questions SAT officielles

Tableau 10 : Résultats de la résolution de GEOS des questions SAT officielles

Résultats de la résolution	Questions répondues correctement	Questions répondues incorrectement	Questions laissées sans réponses
GEOS	27	1	27

- Résultats de la performance de l'extraction :

Tableau 11 : Résultats de la performance de l'extraction

Résultats de la performance de l'extraction	Précision	Rappel	F-mesure
GEOS sans tenir compte des figures	57%	82%	67%
GEOS en tenant compte de la figure	92%	76%	83%

Dans le tableau 11, GEOS est entraîné sur 67 questions qui contiennent un total de 141 relations unaires et 110 relations binaires, et testé sur 55 questions qui contiennent un total de 146 relations unaires et 119 relations binaires.

Pour revenir aux résultats de S^2 [5] du tableau 9, nous pouvons remarquer que la F-mesure de ce dernier qui est 99% est meilleur que celle de GEOS qui est 83%. Cependant, cette comparaison n'est pas très judicieuse, puisque les bases de questions utilisées dans les deux

expérimentations sont différentes par rapport à la langue et aussi par rapport aux niveaux scolaires ciblés.

Dans le tableau 11, on compare ici les résultats de GEOS sans et avec extraction de la figure. GEOS utilise la figure pour les raisons suivantes :

- Les hypothèses extraites du texte sont vérifiées à travers la figure afin de corriger leurs scores obtenus à partir du texte. Prenons en exemple l'énoncé suivant 'La droite AC et perpendiculaire à BD', l'hypothèse extraite de cet énoncé est Perpendiculaire(AC,BD). Cette hypothèse a un score, comme on a expliqué ci-haut, calculé à partir du texte. Cette hypothèse est comparée avec la figure illustrative de cet énoncé en ajoutant aussi son score figural. Par exemple, si dans la figure, l'angle entre AC et BD est proche de 90 degrés alors le score figural est proche de 1 et vice versa. Ce qui explique l'amélioration de la précision dans le tableau 11 au détriment du rappel.
- Pour extraire des *hypothèses visuelles de grand score* : ce sont une liste d'hypothèses qui sont marquées uniquement sur la figure, comme les longueurs des segments et les mesures des angles. Voir la figure 5 (b) où la mesure de l'angle ABC qui est 40 degrés est marquée uniquement dans la figure.

Le résolveur utilisé dans l'expérimentation est conçu par GEOS et il est inspiré des travaux de [18], [19] et [20]. Pour ce faire, on utilise un ensemble d'axiomes constitués des prémisses et des conclusions, illustrés à la figure 8.

Axiom Name	Premises	Conclusion(s)
Midpoint Def.	Midpoint(M, AB)	$AM = MB$
Angle Addition	$\angle ABC, \angle CBD$ Exterior($D, \angle ABC$)	$\angle ABC + \angle CBD = \angle ABD$
Vertical Angles	Intersect(X, AB, CD)	$\angle AXD \cong \angle CXB,$ $\angle AXC \cong \angle BXD$
Side-Side-Side	$\triangle ABC, \triangle DEF,$ $AB \cong DE, BC \cong EF$ $CA \cong FD$	$\triangle ABC \cong \triangle DEF$
Alt. Int. Angles	$CD \parallel EF,$ Intersect(M, AB, CD), Intersect(N, AB, EF)	$\angle ENM \cong \angle NMD,$ $\angle FNM \cong \angle NMC$

Figure 8 : des exemples des axiomes utilisés dans [18]

Grâce à ces axiomes, On génère une succession d'inférences qui permettent d'aller des hypothèses à la conclusion recherchée.

Pour revenir à notre problématique, pour les mêmes raisons que celles évoquées pour S^2 [5], GEOS est un candidat probable pour répondre aux objectifs de notre recherche, puisqu'il a plusieurs avantages :

- GEOS est conçu pour fournir au résolveur les connaissances nécessaires à la résolution, et non pas conçu pour la simplification de la question comme LIM-G.
- Le formalisme d'extraction décrit ci-haut permet d'extraire des hypothèses des énoncés plus complexes où il y a des coréférences entre les phrases, par opposition à S^2 [5]. L'énoncé peut être traité en entier sans avoir besoin de segmenter en phrase.
- La taxonomie, les prédicats et les fonctions peuvent être modifiés et étendus pour prendre en considération les notions géométriques que nous pouvons trouver dans l'école secondaire québécoise.
- GEOS est conçu également pour extraire l'information des figures, ce qui peut être exploitable dans un futur travail.
- De plus, le code source de GEOS avec sa base de données des questions SAT est disponible en accès public.

Toutefois, GEOS a quelques limites :

- GEOS est conçu pour traiter uniquement les questions SAT écrites en anglais. Ce qui signifie que la taxonomie, les prédicats et les fonctions sont plus adaptés à ces questions SAT, par conséquent, ils n'englobent pas toutes les notions géométriques enseignées dans le contexte secondaire québécois qui est plus large (p. ex. dans les questions SAT nous ne trouvons pas la relation d'*intersection entre un cercle et une droite en un point*).
- Comme on a décrit ci-haut, GEOS se restreint à quelques types de conclusion que nous trouvons dans les questions SAT à choix-multiple, et on ne traite pas, par exemple, les questions de démonstration (p.ex. '*ABCD et CDEF sont deux parallélogrammes. Démontre que ABFE est un parallélogramme. Montre que AE est égal à BF*') que nous trouvons dans les manuels scolaires québécois.
- A la différence de S^2 [5] qui utilise uniquement une base de données de patrons pour extraire les informations de l'énoncé, GEOS a besoin d'un grand corpus de questions nécessaire l'entraînement (rappelons qu'on utilise de l'apprentissage machine pour apprendre à GEOS à extraire les prédicats en utilisant l'analyse syntaxique de l'énoncé). On utilise dans GEOS 186 questions SAT à cette fin. De plus, chaque

question d'entraînement doit être annotée avec l'ensemble des hypothèses et des conclusions, ce qui représente une tâche laborieuse, délicate et lente.

En tenant compte des comparaisons faites ci-haut des trois extracteurs de l'état de l'art, nous avons décidé que GEOS est le plus compatible avec les objectifs de notre recherche, malgré ses limites. Dans la section suivante nous décrivons notre stratégie pour adapter l'extracteur GEOS aux problèmes du niveau secondaire québécois.

CHAPITRE 3 MÉTHODOLOGIE

Dans ce chapitre nous présentons l'ensemble des méthodes que nous avons choisies pour vérifier nos hypothèses. Premièrement, nous décrivons la façon dont nous avons procédé pour créer le corpus. Deuxièmement, nous présentons comment nous avons adapté et amélioré la méthode de GEOS pour le type de problèmes que nous devons traiter. Finalement, nous décrivons notre méthode de validation.

3.1. Création du corpus et choix des énoncés de géométrie

3.1.1. Corpus initial

Pour créer notre corpus de problèmes, nous avons d'abord utilisé celui de GEOS [6]. Il s'agit de 186 questions SAT proposées par College Board [16] écrites en anglais. Nous avons traduit ces questions en français. Chaque question est sous forme d'une description textuelle en anglais accompagnée d'une figure. Ce corpus nous a servi comme point de départ de la construction de notre corpus final. Il a l'avantage d'être annoté par GEOS avec les hypothèses et les conclusions que l'extracteur automatique doit trouver. De plus, il est en accès public. Ces questions SAT sont utilisées pour pré-entraîner notre modèle d'extraction.

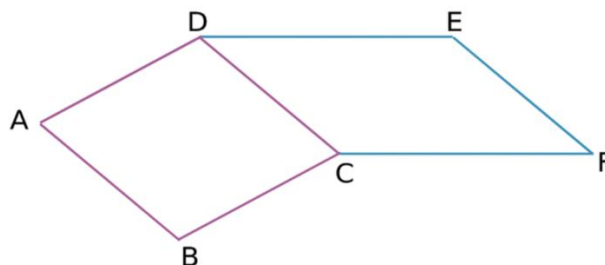
3.1.2. Corpus enrichi

Les questions SAT ne suffisent pas pour apporter un éclairage sur toute la problématique de l'extraction de connaissances à partir d'énoncés de problèmes dans le contexte scolaire québécois. L'extracteur doit être entraîné et testé sur des questions tirées des manuels scolaires et des ressources des professeurs. Ainsi, le corpus est enrichi par l'ensemble de questions supplémentaires suivant:

- Nous avons ajouté des 30 questions tirées des manuels scolaires québécois de première et deuxième années du secondaire. Nous avons ciblé plusieurs manuels [21,22,23,24,25,27,28,29,30,31,32,33,34, 35,36] pour qu'ils couvrent une grande variété de questions. Nous avons choisi uniquement les questions pour lesquelles toute l'information nécessaire à sa résolution est présente dans le texte. Dans

l'exemple ci-dessous, la figure illustrative qui accompagne l'énoncé n'est pas nécessaire à la résolution :

ABCD et CDEF sont deux parallélogrammes. Démontre que ABFE est un parallélogramme. Montre que AE est égal à BF.



- 60 questions choisies par des didacticiens en mathématiques ont été ajoutées pour couvrir le type de questions que nous trouvons dans le matériel didactique des professeurs.

Les deux nouveaux ensembles de questions traitent plus de concepts géométriques (comme polygone, quadrilatère, cercle, ellipse, pentagone, décagone, pentadécagone, etc.) par rapport aux questions SAT. On appelle ces concepts géométriques dans le formalisme de représentation de GEOS des classes géométriques. Voir la taxonomie du formalisme de représentation Ω de la section 2.3.2.4.

Tableau 12 : Comparaison entre le nombre de concepts géométriques présents dans le corpus initial et enrichi

Corpus	Le nombre de concepts
Le corpus initial	34
Le corpus enrichi	94

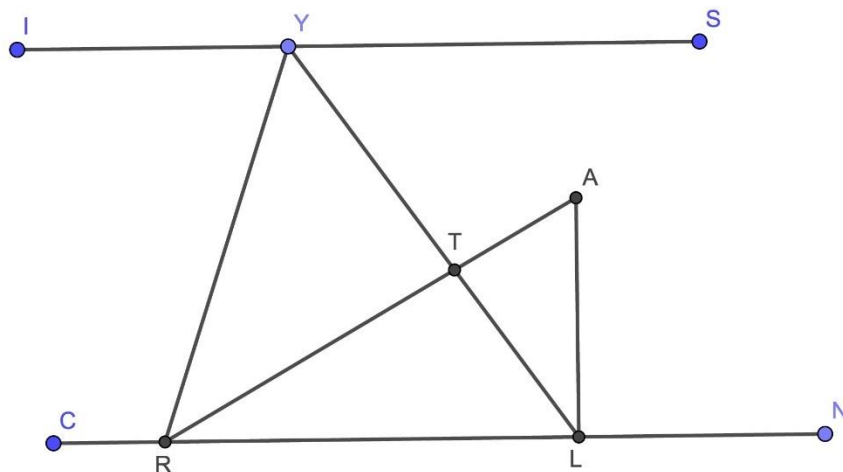
De plus, les questions SAT se caractérisent par des phrases courtes où chaque phrase contient une relation géométrique, alors que dans les deux nouveaux ensembles de questions, nous trouvons des phrases qui sont plus longues et contiennent plus qu'une relation géométrique à la fois. Prenons en exemple les deux questions suivantes :

- Une question tirée des manuels scolaires québécois suivante : Soit $ABCD$ un carré. O est le point d'intersection des diagonales de ce dernier. Soit K , un point de AB , tel que AK égale AO , montrer que la mesure de l'angle AOK est égale à trois fois la mesure de l'angle BOK .
- Une question SAT : $ABCD$ est un parallélogramme. La mesure de BD vaut 2. Les angles du triangle BCD sont tous égaux. Quel est le périmètre p du parallélogramme $ABCD$?

Nous remarquons que les phrases sont plus courtes dans les questions SAT. La deuxième phrase de la question tirée des manuels scolaires ' O est le point d'intersection des diagonales de ce **dernier**' fait une référence au carré ' $ABCD$ ' de la première phrase. Ce genre de coréférence n'est pas beaucoup présent dans les questions SAT.

Ces deux ensembles de questions, plus les questions SAT, sont utilisés à la fois pour l'entraînement et la validation, en effectuant la validation croisée tel que décrit dans la section 3.3. Ces questions ont été annotées manuellement par des didacticiens avec l'ensemble des hypothèses et des conclusions attendues par l'extracteur. Voici un exemple de question annotée :

Soient (IS) et (CN) des droites parallèles. Y est un point de du segment $[IS]$ et R est un point du segment $[CN]$ tels que l'angle IYR mesure 72 degrés. La bissectrice de l'angle RYS coupe $[CN]$ en L . On considère les points T et A tels que T appartient au segment $[YS]$, l'angle ATL mesure 84 degrés et l'angle ALR est droit. Détermine la mesure de l'angle TAL .



Hypothèses :

(IS) est parallèle à (CN)

Y sur [IS]

R sur [CN]

L sur [CN]

T sur [YS]

La bissectrice de l'angle RYS coupe [CN] en L.

ALR est un angle droit

L'angle IYR mesure 72 degrés

L'angle ATL mesure 84 degrés

Mesure ALR=90 degrés

(YL) est la bissectrice de RYS

L'angle RYL

L'angle LYS

YL est une sécante de (IS) et (CN)

YR est une sécante de (IS) et (CN)

Triangle ALR

Triangle ATL

Conclusion :

Mesure de l'angle TAL

3.2. Adaptation de GEOS pour l'extraction des connaissances des énoncés québécois

Dans le chapitre précédent (Revue de littérature), nous avons expliqué les raisons pour lesquelles nous avons choisi GEOS pour l'adapter aux énoncés du secondaire québécois.

Cette adaptation consiste à :

- Réannoter les questions SAT après leurs traductions en français avec les hypothèses et les conclusions. Pour l'entraînement, chaque question est annotée avec les hypothèses et les conclusions qui s'y trouvent avec leurs positions dans le texte. Le passage de l'anglais au français change les positions des mots dans le texte et par conséquent les

positions des hypothèses et les conclusions. Nous expliquons en plus de détails cette réannotation dans le prochain chapitre.

- Adapter les règles d'extraction utilisées dans GEOS pour le français.
- Ajouter les concepts géométriques détectés dans le corpus enrichi dans le formalisme de représentation de GEOS.
- Adapter la sortie de l'extracteur à l'entrée du générateur de preuves de QED-Tutrix.

3.3. Validation de l'extraction

Dans cette section, nous expliquons le processus de validation que nous utilisons pour tester notre extracteur de connaissances. Nous avons opté pour deux critères de validation qui sont le calcul du rappel et de la précision, et au aussi le calcul du pourcentage de questions résolues.

3.3.1 La performance de l'extraction (rappel et précision)

Dans un premier temps, la validation est basée sur le calcul du rappel et de la précision par rapport à l'ensemble des hypothèses et des conclusions extraites manuellement pour chaque question. Cette méthode de validation est la même utilisée dans GEOS [4] et dans S^2 [5]. Le rappel est le rapport du nombre des hypothèses et des conclusions correctement extraites par l'extracteur sur le nombre des hypothèses et des conclusions que l'extracteur est censé extraire. La précision est le rapport du nombre des hypothèses et conclusions correctement extraites par l'extracteur sur le nombre des hypothèses et conclusions extraites par ce dernier.

$$\text{Rappel} = \frac{\sum_{i=1}^N (\text{hypothèses et conclusions correctement extraites})_i}{\sum_{i=1}^N (\text{hypothèses et conclusions})_i}$$

$$\text{Précision} = \frac{\sum_{i=1}^N (\text{hypothèses et conclusions correctement extraites})_i}{\sum_{i=1}^N (\text{hypothèses et conclusions extraites})_i}$$

Où N est le nombre total des questions de la base de données et i est l'indice des questions du corpus. Nous avons opté pour une micro moyenne afin de pouvoir comparer nos résultats avec les résultats de GEOS qui utilisent également la micro moyenne.

À partir de la précision et le rappel, nous calculons la F-mesure qui est une moyenne harmonique de ces deux derniers :

$$F \text{ Mesure} = \frac{2*(rappel*précision)}{rappel+précision}$$

L'algorithme de validation que nous avons utilisé est une *validation croisée à 5 plis* que nous répétons 4 fois tel que décrit ci-bas :

Algorithme de validation : Calculer le rappel, la précision et la F-mesure
Entrée : L'extracteur et l'ensemble des questions EQ du corpus Sortie : le rappel, la précision et la F-mesure Pour $j = 1..4$: Diviser EQ aléatoirement en cinq sous-ensembles $\{EQ_1.., EQ_5\}$; Pour $i = 1..5$ faire : L'ensemble de validation $EV = EQ_i$; L'ensemble d'entraînement $EE = EQ \setminus EQ_i$; Calculer le $rappel_{i,j}$ et la $précision_{i,j}$ pour l'extracteur; Calculer le $rappel_j$ et la $précision_j$ qui sont les moyennes des $rappel_{i,j}$ et des $précision_{i,j}$; Calculer le $rappel$ et la $précision$ qui sont les moyennes des $rappel_j$ et des $précision_j$ pour $j = 1..4$; Calculer le $F \text{ Mesure}$ à partir du $rappel$ et de la $précision$; Retourner le $rappel$, la $précision$ et la $F \text{ Mesure}$

3.3.2. La résolution des problèmes

Dans une seconde étape de validation, les hypothèses et les conclusions extraites des énoncés sont données en entrée au générateur de preuve utilisé dans QED-Tutrix. Il s'agit de savoir si les données extraites sont suffisantes pour permettre au résolveur de trouver au moins une preuve du problème. Dans QED-Tutrix [10] [45], une preuve est constituée d'une succession d'inférences reliant les hypothèses extraites automatiquement à la conclusion recherchée. Une inférence peut être considérée comme une formule atomique composée des prémisses (ABC est un triangle et les angles ABC et ACB sont égaux), d'une justification (un triangle ayant deux angles égaux est isocèle) et des conclusions (le triangle ABC est isocèle en A). La conclusion d'une inférence peut être utilisée comme une prémisse d'une autre inférence, ce qui permet de créer cette chaîne d'inférence reliant les hypothèses à la conclusion du problème. Cette chaîne d'inférence est enregistrée dans un graphe appelé HPDIC. Voir dans la figure 9 l'exemple d'un graphe HPDIC généré pour l'énoncé "Démontrer que le quadrilatère ABCD

ayant les angles BCD, ADC et BAD droits est un rectangle.”. Chaque chemin reliant les hypothèses à la conclusion est une solution possible de la question.

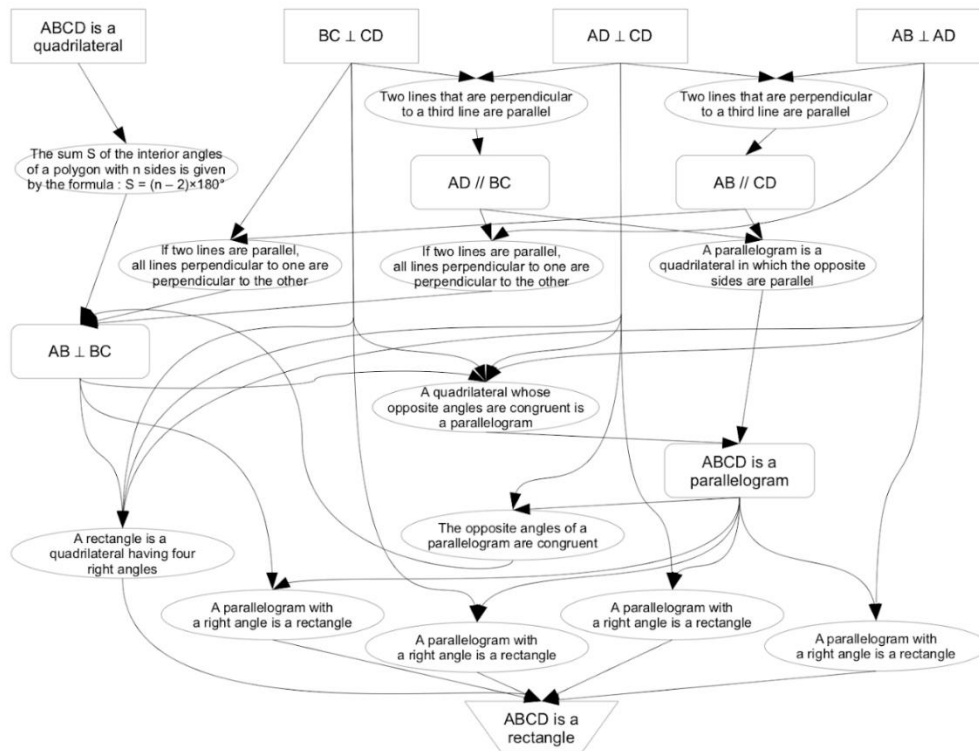


Figure 9 : un exemple d'un graphe HPDIC

Pour chaque question de l'ensemble de tests, nous générons le graphe HPDIC à partir des hypothèses extraites automatiquement. Une fois le graphe HPDIC construit, nous cherchons si la conclusion est présente dans ce dernier pour conclure que la question a été résolue correctement.

L'impact de l'extraction sur la résolution

$$= \frac{\text{Le nombre de questions correctement résolues à partir de l'extraction}}{\text{Le nombre total de questions}}$$

3.3.3. Validation finale

Dans cette section, nous décrivons le test final. 41 questions supplémentaires ont été choisies par des didacticiens pour faire ce test final. Toutes les questions du corpus enrichi décrit dans la section 3.1.2 sont utilisées dans l'entraînement afin de calculer le rappel, la précision et le pourcentage de questions résolues pour ces 41 questions.

Ces questions ne font pas partie de notre corpus enrichi décrit ci-haut. Ces questions sont tirées de plusieurs ressources :

- 9 questions tirées des manuels scolaires québécois de première et deuxième secondaire.
- 15 questions des ressources des professeurs.
- 17 questions prises de la base d'exercices de [57].

Cette liste de questions est donnée dans l'annexe 4.

CHAPITRE 4 ADAPTATION DE LA MÉTHODE GEOS POUR DES PROBLÈMES ÉNONCÉS EN FRANÇAIS

4.1. GEOSFR: adaptation de GEOS aux questions SAT traduites en langue française

4.1.1. Traduction et annotation des questions SAT

La réannotation des questions SAT est nécessaire pour entraîner le classifieur avec les nouvelles positions des hypothèses et des conclusions dans les questions SAT traduites en français.

186 questions SAT ont été traduites en français. À cette fin, nous avons utilisé Google Traduction dans un premier temps, et dans un deuxième temps, cette traduction a été révisée pour éliminer le maximum de fautes de traduction. Ensuite, nous avons annoté ces questions SAT traduites avec les nouvelles positions des concepts identifiés dans la phrase.

Prenons, par exemple, la question suivante de SAT en anglais (voir la figure 10 ci-dessous): “*In the diagram below of triangle ABC, D is a point on the line AB, $AC = 7$, $AD = 6$ and $BC = 18$. Which of the following could be the length of the line DB?*”. Après la traduction en français on obtient :

“*Dans la figure ci-dessous représentant un triangle ABC, D est un point sur la droite AB, $AC = 7$, $AD = 6$ et $BC = 18$. Laquelle des réponses suivantes correspond à la longueur de la ligne DB ?*”.

Les concepts identifiés dans la phrase “*In the diagram below of triangle ABC, D is a point on the line AB, $AC = 7$, $AD = 6$ and $BC = 18$.*” sont les suivants:

Is@9(point@8, point@11)

PointLiesOnLine@11:12(point@11, line@14)

IsLine@14(line@15)

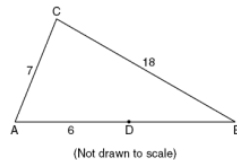
IsTriangle@5(triangle@6)

Les indices @i signifie la position du mot en commençant par 0. Par exemple, le concept Is@9 a été identifié dans la position 9 et le concept point@8 est identifié dans la position 8, etc. (voir la figure 10). Dans le cas de PointLiesOnLine@11:12, ce prédicat est représenté par deux mots ‘point on’, ce qui explique la présence de deux indices.

Le recours à ces indices s’explique par le fait que plusieurs occurrences d’un même mot peuvent se trouver dans le même énoncé. Nous ne pouvons donc pas nous servir uniquement

des mots du texte pour l'entraînement de l'extracteur : Dans l'exemple, l'énoncé 'Le triangle ABC est rectangle en C, l'angle ABC est aigu.', présente deux occurrences du mot ABC, représentant deux concepts différents. Le recours aux indices permet de préciser quelle occurrence correspond à chaque concept.

Question ID: 9



In the diagram below of triangle ABC, D is a point on the line AB, $AC = 7$, $AD = 6$ and $BC = 18$. Which of the following could be the length of the line DB?

sentence index: 0

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
In the diagram below of triangle ABC , D is a point on the line AB , @s_0 holds , @s_1 holds and @s_2 holds .																									

```
Is@9(point@8, point@11)
PointLiesOnLine@12(point@11, line@14)
IsLine@14(line@15)
IsTriangle@5(triangle@6)
```

Parses:

Upload

Next

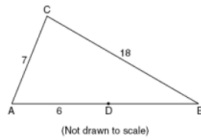
Figure 10 : annotation d'une question SAT originale en anglais avec le formalisme Ω de GEOS.

La nouvelle annotation de cette phrase en français " Dans la figure ci-dessous représentant un triangle ABC, D est un point sur la droite AB, $AC = 7$, $AD = 6$ et $BC = 18$. " donne le résultat suivant (voir la figure 11):

```
Is@12(point@11, point@14)
PointLiesOnLine@15(point@11, line@18)
IsLine@17(line@18)
IsTriangle@8(triangle@9)
```

Nous remarquons que les positions des concepts identifiés dans la phrase ont changé. Par exemple le concept $Is@12$ dans la phrase en anglais a été identifié dans la position 9, ici est la position 11 (voir la figure 11).

Question ID: 9



(Not drawn to scale)
 Dans la figure ci-dessous représentant un triangle ABC, D est un point sur la droite AB, $AC = 7$, $AD = 6$ et $BC = 18$. Laquelle des réponses suivantes correspond à la longueur de la ligne DB ?

sentence index: 0

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 Dans la figure ci - dessous représentant un triangle ABC , D est un point sur la droite AB , @s_0 holds , @s_1 holds et @s_2 holds .

```
Is@12(point@11, point@14)
PointLiesOnLine@15(point@11, line@18)
IsLine@17(line@18)
IsTriangle@8(triangle@9)
```

Parses:

[Next](#)

Upload

Figure 11 : annotation d'une question SAT traduite en français avec la logique de premier ordre de GEOS.

Le changement des positions des concepts entre l'anglais et le français est dû au fait qu'il n'y a pas un alignement exact entre les mots dans les deux langues. Par exemple, le mot "below" a été traduit par trois mots "ci-dessous" (le trait d'union est considéré comme un mot à part dans l'analyse syntaxique par SyntaxNet), ce qui a créé le décalage dans l'exemple ci-dessus.

4.1.2. Adaptation du modèle d'identification des concepts et des relations dans GEOS pour le français (GEOSFR)

Nous détaillons ci-dessous les changements que nous avons introduits pour adapter GEOS au français, nous appelons cette adaptation GEOSFR :

- La première étape de l'identification des concepts dans le texte consiste à étiqueter chaque mot du texte par les concepts de Ω . Cet étiquetage s'effectue de deux façons. D'une part, en utilisant un dictionnaire construit à partir des questions d'entraînement de la base de données. D'autre part, en étiquetant le texte avec un ensemble de règles. Le passage de l'anglais au français a nécessité de changer l'analyseur syntaxique utilisé. Les règles d'étiquetage ont dû être adaptées de la façon suivante :
 - Changement d'étiquettes POS : Les règles d'étiquetage de GEOS se servent de l'étiquetage POS produit par Stanford Parser, alors que nous utilisons SyntaxNet dans l'adaptation pour le français. Cela s'explique par le fait que

Stanford Parser ne produit pas l'arbre de dépendance syntaxique pour le français. Les étiquettes utilisées par les deux analyseurs ne sont pas les mêmes, par exemple un déterminant est désigné par “DT” dans Stanford Parser alors dans SyntaxNet est “DET”. Par conséquent, nous avons dû changer toutes ces règles dans GEOSFR, en remplaçant toutes les étiquettes POS de Stanford Parser par leurs équivalents dans SyntaxNet.

- Traductions de certaines règles : Certaines règles d'étiquetage utilisent des mots en anglais. Par exemple, pour déterminer qu'un point est un cercle, GEOS vérifie que mot précédent est *circle*. Dans ce type de règle, ces mots sont traduits en français.
- Traitement des caractères spéciaux : La présence de certains caractères spéciaux dans la langue française, comme le trait d'union “-”, l'apostrophe “'”, etc. Ces caractères ont des conséquences sur la segmentation de la phrase et aussi sur l'identification des expressions mathématiques. Pour la segmentation de la phrase, GEOS néglige les apostrophes et les traits d'union. Par conséquent, des phrases comme “n'est” seront considérées comme un seul mot, alors que l'analyseur SyntaxNet les considère comme trois mots. Ce qui crée un conflit entre la segmentation faite par GEOS et ce dernier. En ce qui concerne l'identification des expressions mathématiques, le détecteur des expressions mathématiques de GEOS confond le trait d'union avec le trait de soustraction. De ce fait il considère l'expression ‘ci-dessus’ comme une soustraction de “ci” par “dessus”, ce qui est sémantiquement faux.

En ce qui concerne l'extraction des équations mathématiques du type “ $AB + 3 \cdot AC^2 = 82$ ”, nous avons conservé la partie de GEOS dédiée à l'analyse de ce type d'expressions, puisqu'elles sont invariables en fonction de la langue.

4.1.3. Résultats de GEOSFR

Dans le tableau 13, nous comparons les résultats de GEOS sur les questions SAT en anglais avec ceux de GEOSFR sur les mêmes questions traduites en français. Dans cette expérimentation, nous excluons les *hypothèses visuelles de grand score* qui sont extraites uniquement de la figure (voir la section 2.3.2.4) afin que la comparaison soit uniquement sur

les hypothèses du texte. GEOS avec extraction de la figure signifie qu'on utilise les figures pour recalculer les scores des hypothèses déjà extraites du texte.

Tableau 13 : Comparaison des résultats de GEOS sur les questions SAT en anglais avec ceux de GEOSFR sur les mêmes questions traduites en français

Méthodes	Précision (en %)	Rappel (en %)	F-mesure (en %)
GEOS sans extraction de la figure	57	82	67.25
GEOS avec extraction de la figure	92	76	83
GEOSFR	73	65	68.76

Pour avoir les résultats ci-dessus, nous avons utilisé les mêmes ensembles de questions d'apprentissage et de validation utilisés par GEOS. 142 questions sont utilisées pour l'apprentissage du modèle et 44 questions pour sa validation. Nous remarquons que les résultats de GEOSFR sur les questions SAT traduites en français sont très proches en termes de F-mesure (68.76%) de ceux de GEOS sans extraction de la figure sur les questions SAT en anglais 67%. Nous expliquons l'amélioration de la précision par le remplacement du modèle de classification utilisé dans GEOS qui est la *régression logistique* par la *forêt d'arbres décisionnels* dans GEOSFR. Nous expliquons la diminution du rappel, par le fait que GEOS opte pour un rappel plus grand au détriment d'une précision faible. Ce choix se justifie par l'utilisation de l'extraction de la figure dans le GEOS final qui permet de filtrer les faux positifs qui résultent de ce rappel qui est grand. Voir le tableau 13. Cependant, puisque nous n'utilisons pas l'extraction de la figure, nous avons opté pour un rappel et une précision qui soient équilibrés. Ce qui explique l'amélioration de notre F-mesure par rapport à celle de GEOS sans extraction de la figure.

Dans le tableau 14, nous avons appliqué d'autres modèles de classification afin d'améliorer les résultats de l'extraction, à savoir la précision et le rappel. Nous avons remarqué que la Forêt d'arbres décisionnels donne des résultats meilleurs.

Tableau 14 : Résultats de GEOSFR en fonction du modèle de la classification

Modèle de classification	Précision (%)	Rappel (%)	F-mesure (%)
Régression logistique	54	61	57.28

Machine à vecteurs de support	65	61	62.94
Perceptron multicouche	59	62	60.46
K plus proches voisins	50	45	47.37
Forêt d'arbres décisionnels	73	65	68.76
Sans modèle de classification	< 0.1	80	< 0.1

Dans la dernière ligne du tableau 14, nous n'utilisons aucun modèle de classification, c'est-à-dire que nous acceptons toutes les combinaisons des prédicats avec les variables et les constantes sans tenir compte du texte (voir la section 2.3.2.2 : Identification des relations). Nous remarquons que la précision descend à moins de 0.1%, ce qui montre l'importance de l'analyse syntaxique dans l'extraction.

4.2. L'extraction des hypothèses et des conclusions des questions tirées des manuels scolaires québécois ou choisies par les didacticiens

4.2.1. Résultats d'extraction de GEOSFR

Dans la suite, nous appelons les questions tirées des manuels scolaires québécois et celles choisies par les didacticiens le **corpus MSD**.

Pour tester GEOSFR, nous avons utilisé la validation croisée expliquée dans la section 3.3. Le nombre total de questions utilisées est 140. Elles sont composées de :

- 60 questions choisies par des didacticiens et inspirées de ce que nous trouvons dans le matériel didactique utilisé par les enseignants
- 30 questions tirées des manuels scolaires.
- 50 questions SAT parmi le total de 186 questions SAT. Nous avons en choisi uniquement 50 afin que le total des trois ensembles de questions soit équilibré. De plus, nous connaissons déjà la performance de GEOSFR sur les questions SAT, par conséquent, nous ne voulons pas que les résultats soient biaisés par ces dernières.

Nous avons choisi la validation croisée pour calculer les métriques de performance, car elle permet d'utiliser l'intégralité des données pour l'entraînement et pour le test. Ce qui permet de pallier le fait que nous ne disposons pas d'un très grand nombre de questions.

Tableau 15 : Résultats de la validation croisée de GEOSFR sur le corpus MSD

Résultats de la validation croisée de l'extraction	Précision	Rappel	F-mesure
GEOSFR (SAT + questions des didacticiens + questions des manuels)	40%	45%	42.3%
GEOSFR (Questions des didacticiens + Questions des manuels)	30%	28%	28.9%

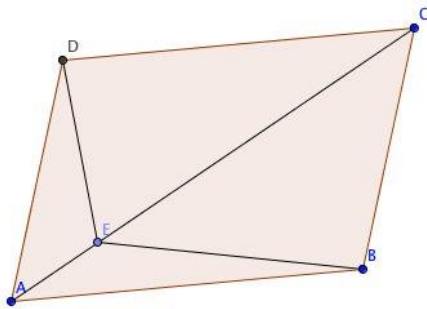
Dans un premier temps, nous avons appliqué la validation croisée sur les 140 questions, ensuite nous l'avons appliquée uniquement sur les questions choisies par les didacticiens et les questions des manuels. Nous ne nous pouvons pas tester sur chaque ensemble séparément parce que le nombre de questions dans chacun est petit. Dans la validation croisée une partie de l'ensemble de questions est utilisée pour l'entraînement et l'autre partie pour la validation. Par exemple, les 30 questions des manuels scolaires ne sont pas suffisantes.

Nous remarquons que les résultats de GEOSFR sur le corpus MSD (en tenant compte des questions SAT ou pas) ne sont pas très bons. Ces résultats s'expliquent par l'insuffisance du formalisme Ω utilisé dans GEOSFR pour représenter les questions du corpus MSD, qui est exactement le même que celui de GEOS. Ce que nous expliquons dans les deux sections suivantes.

4.2.2. Insuffisance du formalisme Ω pour représenter les hypothèses et les conclusions

Dans cette section, nous montrons l'insuffisance du formalisme Ω pour représenter les hypothèses et les conclusions nécessaires à la résolution des problèmes du corpus MSD. Pour ce faire, nous donnons des exemples illustrant les lacunes de Ω . Cette liste d'exemples n'est pas exhaustive, cependant elle couvre la majorité des problèmes qui sont surtout liés aux conclusions. Cela s'explique par le fait que la grande différence entre les questions SAT à choix multiples et les questions du corpus MSD est le type des conclusions qui s'y trouvent.

Exemple 1 : “Si E est un point quelconque de la diagonale AC du parallélogramme ABCD, quelle relation y a-t-il entre les aires des triangles AEB et AED ?”



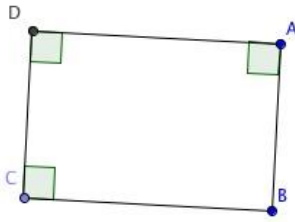
Dans cet exemple, l'ensemble des hypothèses que nous pouvons formuler avec le formalisme Ω est le suivant:

- $Is(E, point)$
- $PointLiesOnLine(E, AC)$
- $IsTriangle(AEB)$
- $IsTriangle(AED)$

Cet ensemble ne représente pas toutes les informations présentes dans le texte :

- Il manque l'hypothèse “AC est une diagonale de ABCD”, car la liste des prédicats ne contient pas une relation entre la diagonale et le quadrilatère.
- Il manque également la conclusion “quelle relation y a-t-il entre les aires des triangles AEB et AED ?”. Il manque dans le formalisme Ω un prédicat qui représente la relation *rapport(nombre, nombre)*.

Exemple 2 : “Démontrer que le quadrilatère ABCD ayant 3 angles droits est un rectangle.”



Dans cet exemple, le formalisme Ω ne permet de formuler que la seule hypothèse suivante :

- $\text{IsQuad}(ABCD)$

Il manque à cet ensemble les hypothèses implicites :

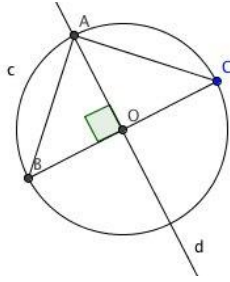
- $\text{IsRightAngle}(BCD)$
- $\text{IsRightAngle}(CDB)$
- $\text{IsRightAngle}(DAB)$

Remarquons que ce ne sont pas les seuls angles que nous pouvons choisir comme angles droits, parce que l'énoncé ne précise pas quels sont les angles droits.

Cet exemple soulève deux lacunes :

- En général, GEOS est incapable avec son mécanisme d'extraction de comprendre des expressions comme “une entité ayant n variables”, où l'extracteur est amené à considérer n variables quelconques de l'entité géométrique en question.
- La conclusion de l'exemple 2 “Démontrer que ... ABCD ... est un rectangle” ne rentre dans aucune des catégories de conclusions traitées par GEOS (**Which**, **What** et **Find**). Il manque à Ω un prédicat *Conclusion(hypothèse)*. Ainsi, nous précisons au générateur de preuve que cette hypothèse est une conclusion que nous cherchons à démontrer. Dans l'exemple 2, la conclusion sera *Conclusion(IsRectangle(ABCD))*

Exemple 3 : “De quelle nature est le triangle ABC si nous avons un cercle C de centre O et de diamètre BC, d la médiatrice du segment BC et A un point d'intersection du cercle et de d.”



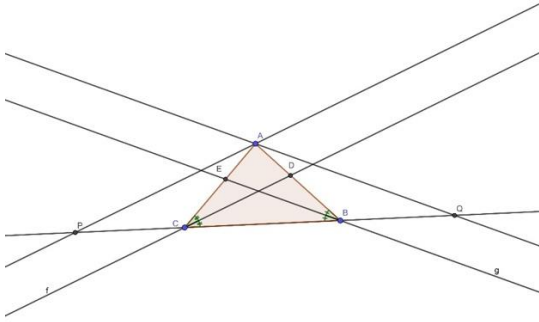
Dans cet exemple, l'ensemble des hypothèses que nous pouvons formuler avec le formalisme Ω est :

- $\text{IsDiameterLineOf}(BC, c)$
- $\text{IsPoint}(A)$
- $\text{IsTriangle}(ABC)$
- $\text{IsCircle}(c)$
- $\text{IsCenter}(O, c)$

Cet exemple soulève trois lacunes :

- Nous ne pouvons pas formuler les hypothèses suivantes :
 - “A est un point d'intersection du cercle et de d”, car le prédicat $\text{IntersectAt}(\text{line}, \text{line}, \text{point})$ de Ω ne considère que l'intersection entre des lignes, ici nous avons l'intersection entre un cercle et une droite.
 - “d est la médiatrice de BC” puisqu'il n'existe pas un prédicat $\text{Médiatrice}(\text{ligne}, \text{ligne})$ pour formuler cette hypothèse.
- Nous ne pouvons formuler la conclusion “trouver la nature du triangle ABC”. Pour combler cette lacune, nous proposons un prédicat $\text{TrouverNature}(\text{entité})$ qui informe le générateur de preuves que nous cherchons à trouver la nature de l'entité en question. Dans l'exemple 3, la conclusion sera $\text{TrouverNature}(ABC)$.

Exemple 4 : “Soit ABC un triangle quelconque, on appelle g la bissectrice issue du sommet B et f la bissectrice issue du sommet C. La droite g coupe AC en E et la droite f coupe AB en D. La droite parallèle à f et passant par A coupe BC en P et la parallèle à g passant par A coupe BC en Q. Est-ce que la longueur du segment PQ est égale au contour du triangle ABC ? Si oui, montre pourquoi ?”

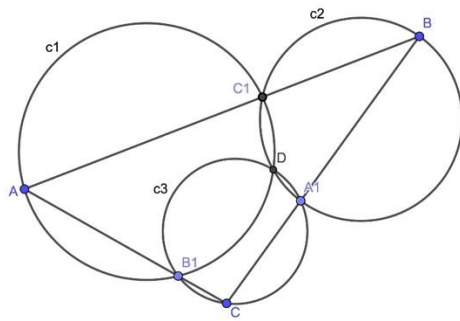


Dans cet exemple, le formalisme Ω ne permet pas de formuler les hypothèses implicites suivantes:

- $\text{BisectsAngle}(g, ABC)$ qui représente l'expression "g la bissectrice issue du sommet B", ici 'g' est la bissectrice de l'angle 'ABC' que nous déduisons implicitement par la considération du triangle ABC.
- $\text{BisectsAngle}(f, BCA)$ qui représente l'expression "f la bissectrice issue du sommet C". Pour la même raison que l'hypothèse d'avant, GEOS est incapable d'extraire cette hypothèse implicite.

La conclusion "Est ce que la longueur du segment PQ est égale au contour du triangle ABC ? Si oui, montre pourquoi." ne rentre dans aucun type de conclusion traitée par Ω . Il s'agit ici d'une question fermée à propos de l'hypothèse "la longueur du segment PQ est égale au contour du triangle ABC", suivie par la requête d'une démonstration de cette hypothèse dans le cas où il est vrai. Pour formuler cette conclusion, nous proposons le prédicat $\text{Conclusion}(\text{hypothèse})$ qui précise au générateur de preuves qu'il s'agit d'une conclusion que nous cherchons à trouver. La conclusion sera $\text{Conclusion}(\text{Equal}(\text{LengthOf}(PQ), \text{PerimeterOf}(ABC)))$.

Exemple 5 : Soit ABC un triangle et A1, B1 et C1 trois points respectivement sur ses côtés [BC], [CA] et [AB]. Montrer que les cercles circonscrits c1, c2 et c3 aux triangles AB1C1, BA1C1 et CB1A1 se rencontrent en un point D.



Comme nous avons vu dans l'exemple 2, avec un prédicat comme *Conclusion(hypothèse)* nous pouvons représenter les conclusions de démonstration. Cependant, l'*hypothèse* du prédicat *Conclusion(hypothèse)* ne peut pas être formulée avec le formalisme Ω , car :

- Premièrement, “les cercles circonscrits aux triangles AFG, BGE et CEF” ne peut pas être exprimé par le formalisme Ω puisqu’il ne connaît pas le prédicat *circonscrit*.
- Deuxièmement, “les cercles ... se rencontrent en un point D” ne peut être exprimée dans Ω puisqu’il manque le prédicat *IntersectIn(point, cercle, cercle)*.

En analysant le formalisme Ω , nous avons remarqué que celui-ci est très biaisée par la forme des questions SAT. En effet, il ne contient que les concepts qui ont été vus dans les questions SAT, ce qui est insuffisant pour couvrir tous les concepts enseignés dans le niveau secondaire québécois. Pour combler ce manque, nous avons choisi une collection de manuels scolaires de mathématiques québécois pour enrichir le formalisme Ω avec plus de concepts. Ce travail a été fait en collaboration avec un étudiant en maîtrise de didactique qui fait partie de l'équipe du laboratoire de recherche Turing. Les manuels scolaires choisis sont Visions, SOMMETS, Sésamath, Point de vue, Point de mire, Perspective, Panoram@th, Mathématique 3000 et Intersection. Les niveaux scolaires ciblés sont la première et la deuxième secondaire pour les filières Culture, société et technique (CTS), Technico-sciences (TS) et Sciences naturelles (SN).

Des canevas (voir la figure 12) pour l'extraction des hypothèses et des conclusions ont été donnés à des étudiants du baccalauréat de deux universités (UdeM et UQÀM), avec une sélection de questions de géométrie présentes dans ces manuels. Une fois ces canevas remplis par ces étudiants, ils ont été utilisés, d'une part pour identifier les concepts qui ne sont pas présents dans les questions SAT (voir le tableau 16), d'autre part pour identifier et analyser les

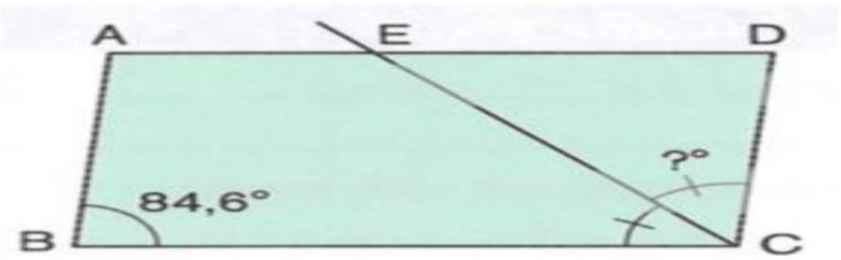
différents types d'hypothèses et de conclusions présentes dans ces questions. Dans la figure 12, les étudiants devaient extraire les hypothèses textuelles, figurales et auxiliaires, plus les conclusions. Les hypothèses auxiliaires sont les hypothèses liées à la résolution où l'étudiant ajoute des nouvelles entités géométriques durant la résolution, qui ne sont pas présentes ni dans le texte ni dans le diagramme.

Feuille 1 (à l'ordinateur)

Manuel : Sommets Chapitre : 3.3 N°6 p. 157

Énoncé

La figure ABCD est un parallélogramme et CE est la bissectrice de l'angle C. Trouve la mesure de l'angle DCE.



Hypothèse textuelles :

- ABCD est un parallélogramme.
- CE est la bissectrice de l'angle C.

Hypothèses figurales :

- $AB \parallel DC$.
- $m(AB) = m(DC)$.
- $AD \parallel BC$.
- $m(AD) = m(BC)$.
- $m\angle ABC = 84.6^\circ$.
- E est un point du segment AD
- $m\angle BCE = m\angle ECD$

Hypothèse auxiliaire :

Conclusion :

Mesure de l'angle DCE.

Figure 12 : exemple d'un canevas avec un exercice tiré du manuel Sommet

Tableau 16 : les concepts identifiés dans les canevas des exercices tirés des manuels scolaires

concepts
polygone, quadrilatère, cercle, demi-cercle, ellipse, pentagone, décagone, pentadécagone, hexagone, octogone, heptagone, dodécagone, hendécagone, triangle, cerf-volant, trapèze, losange, rectangle, carré, arc, segment, rayon, corde, côté, diagonale, hypoténuse, cathète, diamètre, hauteur, médiane, droite, médiatrice, bissectrice, axe, tangente, sécante, point, foyer, milieu, sommet, orthocentre, centre, centre de gravité, intersection, abscisse, ordonnée, vertical, horizontal, focal, nombre, mesure, moyenne, distance, périmètre, rapport, circonférence, longueur, largeur, aire, transformation, similitude, agrandissement, homothétie, isométrie, réflexion, translation, rotation, est médiatrice de, est sécante de, homologue, parallèle, perpendiculaire, adjacent, opposé, congrue, est Médiane de, est rayon de, est axe de symétrie, est côté de, est diagonal de, et hypoténuse de, est cathète de, est

hauteur de, est inscrit dans, est isométrique, est semblable, Image par transformation, est périmètre de, appartient à, est milieu de, est extrémité de, appartient à, intérieur, extérieur, périmètre, adjacent, alternes externes, alternes internes, complémentaire, consécutif, correspondant, opposé par sommet, supplémentaire, coupe en, tangente en

Ces concepts ont été combinés avec ceux du formalisme Ω plus ceux identifiés dans les questions choisies par les didacticiens afin de couvrir le maximum possible de concepts enseignés dans l'école secondaire québécoise. Le formalisme étendu ainsi obtenu a permis de développer GEOSFR++, un extracteur beaucoup plus performant, qui est le sujet du prochain chapitre.

CHAPITRE 5 CRÉATION DE L'EXTRACTEUR GEOSFR++

En combinant les concepts du formalisme Ω avec les concepts identifiés dans le corpus MSD, nous avons créé un nouveau formalisme plus étendu, dénommé Ω_+ , qui couvre le maximum possible des concepts enseignés dans l'école secondaire québécoise.

5.1. Les schémas de la hiérarchie des classes (taxonomie)

La taxonomie de Ω_+ contient 94 classes dont la structure hiérarchique est la suivante:

root

booléen

cc

truth

nombre

mesure

moyennée

distance

périmètre

deSimilitude

dHomothétie

circonférence

longueur

largeur

aire

rapport

entité

point

foyer

milieu

sommet

orthocentre

centre

centreDeGravité

intersection

ligneOuvrte

droite

médiatrice

bissectrice

axe

deRéflexion

desAbscisses

desOrdonnées

deSymétrie

vertical

horizontal

focal

petitAxe

- grandAxe
 - tangente
 - sécante
- segment
 - médiane
 - hauteur
 - diamètre
 - cathète
 - hypoténuse
 - diagonale
 - côté
 - corde
 - rayon
- courbe
 - arc
- ligneFermée
 - cercle
 - demiCercle
 - polygone
 - polygoneDOrdrePair
 - quadrilatère
 - cerfVolant
 - parallélogramme
 - carré
 - rectangle
 - losange
 - trapèze
 - dodécagone
 - décagone
 - hexagone
 - octogone
 - polygoneDOrdreImpair
 - pentadécagone
 - pentagone
 - hendécagone
 - triangle
 - scalène
 - isocèle
 - équilatéral
 - isoangle
 - acutangle
 - équiangle
 - obtusangle
 - rectangle
 - heptagone
 - ellipse
 - étoile
 - angle
 - aigu
 - nul

```

droit
obtus
plat
plein
surface
  région
  secteurAngulaire
  secteurCirculaire
transformation
  similitude
    agrandissement
    isométrie
      réflexion
      rotation
        horaire
        antihoraire
      translation
    similitude
    homothétie

```

Dans l'annexe 1, nous présentons une version graphique de cette hiérarchie. Cette taxonomie a été révisé et corrigé par des didacticiens. Le nombre de classes dans Ω est 34. Il y a donc 60 classes de plus dans Ω_+ , ce qui représente une augmentation substantielle. Nous n'avons pas seulement ajouté des nouvelles classes, nous avons aussi modifié la répartition des classes, c'est-à-dire les relations de sous-type entre les classes. Cette nouvelle répartition est plus cohérente. Une des inconsistances que nous avons trouvées dans Ω est la confusion entre les concepts de segment et de droite. Nous pouvons calculer la longueur d'un segment, mais nous ne pouvons pas le faire pour une droite. De ce fait, nous séparons dans Ω_+ les deux concepts.

Nous avons créé cinq classes disjointes qui catégorisent toutes les entités de la géométrie plane que nous avons identifiées dans notre corpus : *angle*, *point*, *ligne ouverte*, *ligne fermée* et *surface*. Par rapport à Ω , la nouvelle taxonomie contient deux nouvelles classes: *transformation* et *surface*. Ces deux classes ne sont pas présentes dans les questions SAT. La classe *transformation* qui est une sous-classe de la racine *root* contient les transformations géométriques comme l'isométrie, la rotation, translation, etc. Nous avons aussi ajouté une nouvelle classe *surface*, sous-classe *d'entité*, qui était confondue avec la classe aire dans Ω . Ce qui n'était pas pertinent puisque nous pouvons parler d'une surface sans avoir besoin de calculer sans aire. La classe *surface* a trois sous-classes *secteurCirculaire*, *secteurAngulaire*, et *région*.

5.2. La liste des prédicats et des fonctions de Ω_+

Dans le tableau de l'annexe 3, nous présentons la liste des prédicats et des fonctions que nous utilisons pour décrire les différentes relations géométriques possibles entre les classes définies ci-haut. Le nombre de fonctions et de prédicats est de 44 dans Ω et 220 dans Ω_+ . Nous avons repris quelques fonctions et prédicats de Ω . Toutefois, la majorité de ses prédicats et ses fonctions sont redéfinies conformément avec la nouvelle taxonomie de Ω_+ . Nous avons ajouté de nouveaux prédicats et fonctions pour couvrir les concepts qui ne sont pas présents dans les questions SAT. En ce qui concerne les problèmes relatifs aux conclusions que nous avons évoqués pour les questions du corpus MSD, nous avons créé de nouveaux prédicats et fonctions que nous pouvons voir en gras dans l'annexe 3 (p. ex. **Conclusion(hypothèse)->booléen**, **TrouverNature (entité)->booléen**, etc.).

Dans les sections suivantes, nous allons revenir sur quelques prédicats et fonctions qui sont les plus importants, en expliquant leur fonctionnement.

5.3. Description de l'algorithme GEOSFR++ de l'extraction basée sur la Ω_+ avec des exemples

Dans cette section nous présentons un nouvel algorithme d'extraction GEOSFR++ basé sur le formalisme Ω_+ . Nous utilisons pour décrire cet algorithme l'exemple illustratif suivant :

“Si E est un point quelconque de la diagonale AC du parallélogramme ABCD, quelle relation y a-t-il entre les aires des triangles AEB et AED ?”

Il s'agit d'une question du corpus MSD présentée à la section 4.2.2, où nous avons montré l'insuffisance du formalisme Ω pour représenter les hypothèses et les conclusions. Nous supposons dans cette description que le classifieur est déjà entraîné et que l'exemple illustratif est une question de l'ensemble de validation. Puisque la construction du vecteur caractéristique est la même dans l'entraînement et dans la validation, alors, nous nous restreignons à présenter uniquement la partie de validation.

Dans la figure 13, nous présentons une vue d'ensemble de l'algorithme GEOSFR++. Chaque partie de cet algorithme est expliquée dans les sections qui viennent.

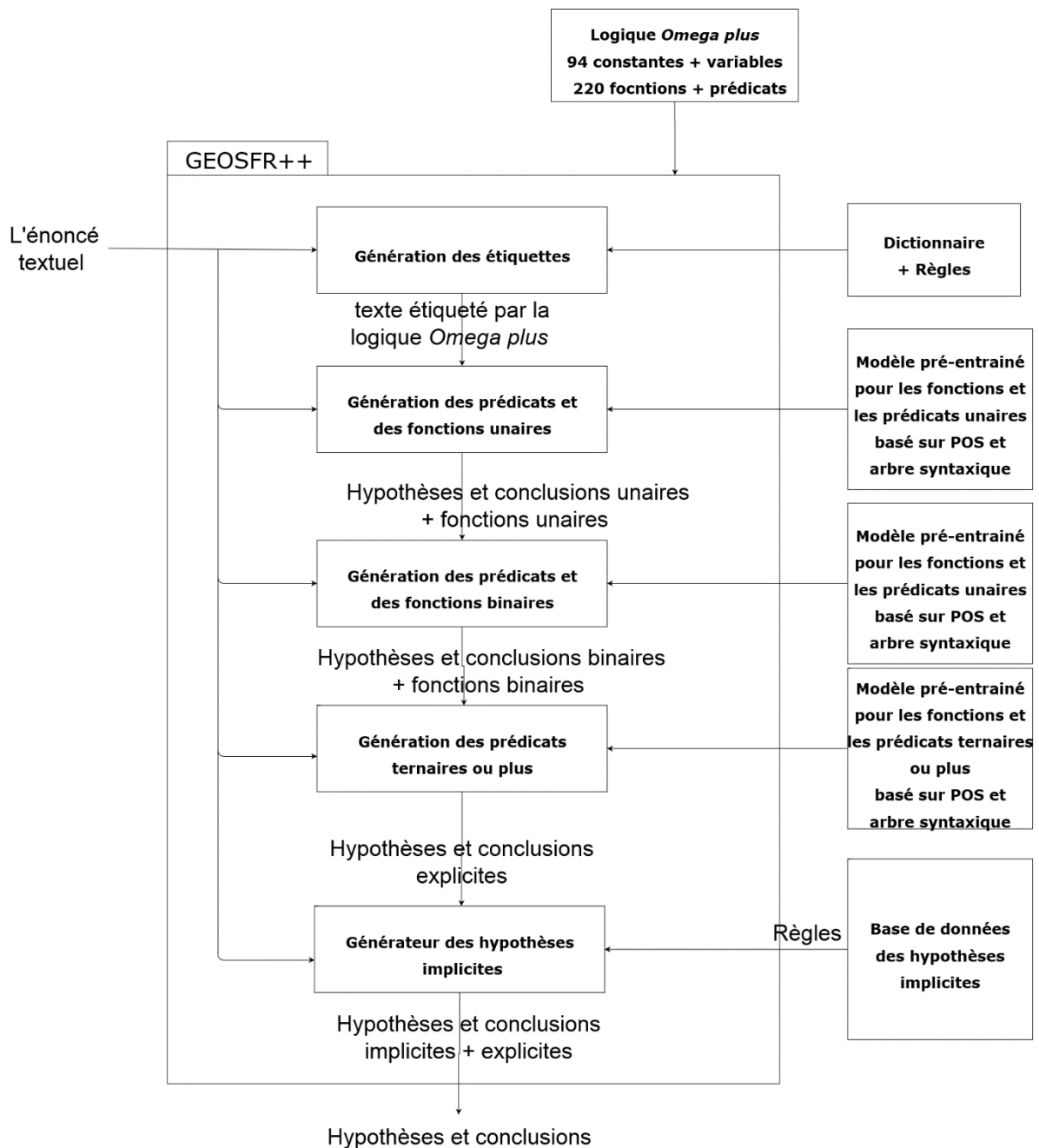


Figure 13 : vue d'ensemble de l'algorithme GEOSFR++

5.3.1. Génération des étiquettes

Nous utilisons les concepts (les classes, les prédicats et les fonctions) du formalisme Ω_+ pour étiqueter l'énoncé du problème. Une liste de règles et un dictionnaire sont utilisés pour faire l'étiquetage. Nous utilisons le même processus d'étiquetage que celui de GEOSFR.

Notons qu'un mot du texte peut avoir une ou plusieurs étiquettes (p. ex. le mot 'ABC' peut être étiqueté comme classe 'angle' ou 'triangle').

Le dictionnaire d'étiquetage consiste à associer à chaque concept de Ω_+ un ensemble de mots que nous trouvons dans l'énoncé. Par exemple, si dans l'ensemble d'entraînement nous rencontrons le mot 'médiatrices' qui est associé au prédicat *EstMédiatriceDe*, alors nous ajoutons dans le dictionnaire le couple ('médiatrices', *EstMédiatriceDe*). Par conséquent, la prochaine fois que nous rencontrons le mot 'médiatrices' dans la validation il sera étiqueté *EstMédiatriceDe*. Pour prendre en compte les variations en genre et en nombre, nous ajoutons également les lemmes des mots dans ce dictionnaire. Ainsi, nous ne sommes pas obligés d'entraîner notre modèle sur toutes les variations en genre et en nombre du même concept. Il suffit de l'avoir vu une fois dans l'entraînement qu'il soit au singulier ou pas ou au masculin ou pas pour être pris en compte. Cette amélioration est née du constat selon lequel, dans GEOS, des mots, par exemple, en singulier, n'ont pas été étiquetés parce que nous avons vu uniquement leurs pluriels dans l'entraînement. Il en va du même pour le genre. Dans l'exemple ci-haut, le lemme de 'médiatrices' qui est 'médiatrice' est aussi associé à *EstMédiatriceDe*.

Prenons l'exemple suivant :

Si E est un point quelconque de la diagonale AC du parallélogramme ABCD, quelle relation y a-t-il entre les aires des triangles AEB et AED?

Nous avons à la sortie l'étiquetage suivant :

Si **E**(point / cercle / demi-cercle / ellipse / ligne / surface) est un **point**(*EstPoint* / point) quelconque de la **diagonale**(*EstDiagonale* / *EstDiagonaleDe* / diagonale) **AC**(droite / segment / arc) du **parallélogramme**(*EstParallélogramme* / parallélogramme) **ABCD**(quadrilatère), **quelle**(Quel) **relation**(*EstRapport*) y a-t-il entre les **aires**(*EstAireDe* / aire) des **triangles** (*EstTriangle* / triangle) **AEB**(angle / triangle / surface) **et**(CC) **AED**(angle / triangle / surface)?

Seuls les mots en gras sont étiquetés. Les mots en italique sont les concepts (les variables, constantes, les prédicats et les fonctions) du formalisme Ω_+ .

5.3.2. Génération des prédicats et des fonctions unaires

Les prédicats et les fonctions sont combinés à toutes les variables et les constantes qui correspondent aux classes des arguments. C'est-à-dire, si par exemple nous avons le prédicat *EstAngle(angle)*, ce dernier ne sera associé qu'aux variables et constantes qui sont étiquetées *angle*. Supposons par exemple que dans un énoncé nous avons l'expression "l'angle ABC", qui a pour étiquetage "**l'angle**(*EstAngle*) **ABC**(*triangle/angle*)". Le prédicat *EstAngle(angle)* est associé à l'angle **ABC**. Le triangle **ABC** sera éliminé parce qu'il n'existe aucun prédicat acceptant un triangle comme argument qui peut être combiné avec lui. Par conséquent, nous aurons à la sortie *EstAngle(ABC)*.

Prenons un deuxième exemple qui contient plusieurs prédicats et variables dans le même énoncé : "L'angle ABC est aigu et l'angle CBD est droit" a pour étiquetage "**l'angle**(*EstAngle/angle*) **ABC**(*triangle/angle*) est **aigu**(*EstAigu*) et **l'angle**(*EstAngle/angle*) **CBD**(*triangle/angle*) est **droit**(*EstDroit*)". Les combinaisons possibles entre les prédicats et les variables dans cette phrase sont {*EstAngle(ABC)*, *EstAngle(CBD)*, *EstAigu(ABC)*, *EstAigu(CBD)*, *EstDroit(ABC)*, *EstDroit(CBD)*}. Nous remarquons que cet ensemble contient des combinaisons correctes, qui sont {*EstAngle(ABC)*, *EstAngle(CBD)*, *EstAigu(ABC)*, *EstDroit(CBD)*} et des combinaisons fausses, qui sont {*EstAigu(CBD)*, *EstDroit(ABC)*}. Comme dans GEOS et GEOSFR, un classifieur déjà entraîné (machine learning) est utilisé pour déterminer si une combinaison est correcte ou fausse.

Le vecteur caractéristique de chaque combinaison est construit en encodage à chaud (one-hot), comme GEOS, à partir de :

- La relation syntaxique entre le prédicat (ou la fonction) et son argument. La figure 14 présente l'analyse syntaxique de l'exemple précédent 'L'angle ABC est aigu et l'angle CBD est droit'. Dans la relation *EstAngle(ABC)*, le prédicat *EstAngle* est relié à son argument par la relation APPOS (voir la figure 14). GEOS prend en considération uniquement les relations syntaxiques que lorsque la distance dans l'arbre de dépendance est inférieure ou égale à deux entre les prédicats et les arguments. Dans la nouvelle version GEOSFR++ nous ne nous limitons pas à une distance de deux, mais nous ajoutons toutes les relations syntaxiques entre le prédicat et ses arguments. Cette modification permet d'améliorer nos résultats de performance, au détriment d'un vecteur caractéristique plus grand. Prenons en exemple le prédicat *EstAigu(ABC)*, la distance syntaxique entre le prédicat et l'argument est 3, ainsi le vecteur caractéristique

dans GEOS ne va pas contenir les relations syntaxiques entre eux, ce qui n'est pas le cas avec GEOSFR++.

- La distance entre le prédicat (ou la fonction) et son argument dans l'énoncé : Nous calculons le nombre de mots entre le prédicat et son argument en ajoutant 1. C'est aussi la différence entre les indices des mots (l'énoncé est indexé en commençant par 0). Prenant le même exemple de la figure 14, la distance pour la relation *EstAngle(ABC)* entre le prédicat et l'argument dans l'énoncé est 1, parce qu'il n'existe aucun mot entre eux : 'angle ABC'.
- La distance entre le prédicat (ou la fonction) et son argument dans le graphe de l'analyse syntaxique : Nous calculons le plus court chemin qui relie le prédicat et son argument dans le graphe de l'analyse syntaxique. Prenant le même exemple de la figure 14, la distance dans le graphe de l'analyse syntaxique pour la relation *Estdroit(CBD)* entre le prédicat et l'argument est égale à 3, parce qu'ils sont reliés par trois relations APPOS, NSUB et ATTR : voir la figure 14.
- La nature grammaticale (étiquettes POS) du prédicat (ou la fonction) et l'argument. Dans la figure 14, dans la relation *EstAngle(ABC)*, *EstAngle* est étiqueté NOUN et ABC est étiqueté NOUN.

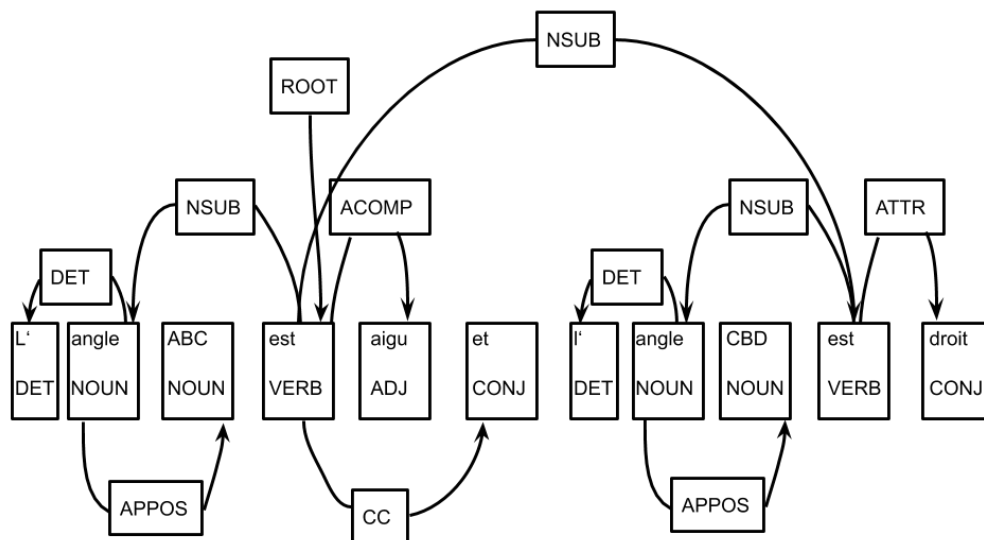


Figure 14 : L'analyse syntaxique de l'énoncé 'L'angle ABC est aigu et l'angle CBD est droit' obtenue par SyntaxNet

Notons que les fonctions peuvent être utilisées comme un argument d'entrée pour les prédicats unaires. Si c'est le cas, elles sont combinées à ces prédicats unaires. Sinon, elles sont laissées pour les phases qui suivent puisqu'il y a une possibilité qu'elles appartiennent à des prédicats binaires ou ternaires.

Soit notre exemple, que nous répétons ici :

Si **E**(*point / cercle / demi-cercle / ellipse / ligne / surface*) est un **point**(*EstPoint / point*) quelconque de la **diagonale**(*EstDiagonale / EstDiagonaleDe / diagonale*) **AC**(*droite / segment / arc*) du **parallélogramme**(*EstParallélogramme / parallélogramme*) **ABCD**(*quadrilatère*), **quelle**(*Quel*) **relation**(*EstRapportEntre*) y a-t-il entre les **aires**(*EstAireDe / aire / AireDe*) des **triangles** (*EstTriangle / triangle*) **AEB**(*angle / triangle / surface*) **et**(*CC*) **AED**(*angle / triangle / surface*)?

À la sortie nous avons :

- *EstPoint(E)*
- *EstDiagonale(AC)*
- *EstParallélogramme(ABCD)*
- *EstTriangle(AEB)*
- *AireDe(AEB)->aire*

5.3.3. Génération des prédicats et des fonctions binaires

Nous avons séparé l'extraction unaire et l'extraction binaire parce que la construction du vecteur caractéristique pour cette dernière prend en compte deux arguments au lieu d'un seul. Mis à part cela, les prédicats et les fonctions avec deux arguments seront générés (c'est-à-dire, le processus de combinaison des variables et des constantes avec les prédicats et les fonctions) et classifiés de la même façon que les prédicats et les fonctions unaires. De même, le vecteur caractéristique des prédicats binaires sera construit à partir des relations syntaxiques, des distances et des natures grammaticales de la propriété et de ses deux arguments. En ce qui concerne les relations syntaxiques, nous adoptons l'amélioration décrite ci-haut, où les distances syntaxiques supérieures à deux sont aussi prises en compte.

Prenant par exemple l'énoncé suivant : "L'angle ABC est aigu et l'angle CBD est droit, ces deux angles sont adjacents. L'angle EFJ est droit.". D'après les étapes précédentes, nous avons les hypothèses unaires suivantes : $\{EstAngle(ABC), EstAngle(CBD), EstAigu(ABC), EstDroit(CBD), EstDroit(EFJ)\}$. Le processus de combinaison des prédicats binaires avec leurs arguments donne : $\{AnglesAdjacents(ABC,CBD), AnglesAdjacents(CBD,EFJ), AnglesAdjacents(ABC,EFJ)\}$. Grâce à notre modèle de classification, nous pouvons déterminer la seule hypothèse binaire vraie qui est $\{AnglesAdjacents(ABC,CBD)\}$.

Reprenant notre exemple, à l'entrée, nous avons, après avoir extrait les prédicats et fonctions unaires :

Si **E**(point / cercle / demi-cercle / ellipse / ligne / surface) est un **point**(EstPoint / point) quelconque de la **diagonale**(EstDiagonale / EstDiagonaleDe / diagonale) **AC**(droite / segment / arc) du **parallélogramme**(EstParallélogramme / parallélogramme) **ABCD**(quadrilatère), **quelle**(Quel) **relation**(EstRapportEntre) y a-t-il entre les **aires**(EstAireDe / aire) des **triangles** (EstTriangle / triangle) **AEB**(angle / triangle / surface) **et(CC)** **AED**(angle / triangle / surface)?

- *EstPoint(E)*
- *EstDiagonale(AC)*
- *EstParallélogramme(ABCD)*
- *EstTriangle(AEB)*
- *AireDe(AEB)->aire*

À la sortie du module d'extraction des prédicats et fonctions binaires, nous aurons les ajouts suivants:

- *EstDiagonaleDe(AC,ABCD)*
- *CC(AEB,AED) implique AireDe(AED)->aire et EstTriangle(AED)*
- *Rapport(AireDe(AED),AireDe(AED)->rapport*

La conjonction de coordination *CC(AEB,AED)* permet d'impliquer deux relations sur le triangle **AED** : *EstTriangle(AED)* et *AireDe(AED)->aire*.

5.3.4. Génération des prédicats et des fonctions ternaires

Pour couvrir les limites de GEOS et GEOSFR, concernant certains types de relations qui ne peuvent pas être exprimées par les relations binaires ou unaires, il était nécessaire d'introduire un nouveau type d'hypothèses ternaires. Le traitement des relations ternaires ou plus dans GEOS et GEOSFR est contourné par un processus impliquant l'utilisation des relations binaires et des conjonctions de coordination. En fait, GEOS et GEOSFR contiennent trois relations qui ne sont ni unaires ni binaires : *IntersectAt(*line,point)*, *AverageOf(*number)->number* et *SumOf(*number)->number*. **line* signifie que *IntersectAt* peut prendre plusieurs lignes en argument. De même pour les deux autres fonctions. Par exemple, pour représenter l'expression "l et d se coupent en E", GEOS (et GEOSFR) a besoin de passer par deux représentations intermédiaires *CC(d,l)* et *IntersectAt(d,E)*. Par la suite, GEOS impliquera l'hypothèse *IntersectAt({d,l},E)*. Ce processus est très limité, puisqu'il est basé sur le fait que *d* et *l* sont de la même classe *line* et qu'elles sont liées par une conjonction de coordination. En fait, ce processus ne permet pas de traiter les entités géométriques de différentes classes (par exemple, l'intersection d'un cercle avec une droite dans un point) ou dans le cas où les entités ne sont pas en conjonction de coordination (Par exemple l'expression "d coupe l en E", dans cette expression *d* et *l* ne sont pas en conjonction de coordination).

Dans GEOSFR++, nous avons introduit un troisième type de prédicats et fonctions qui sont ternaires. Nous avons créé 29 prédicats et fonctions de ce type. Voici quelques exemples :

- *EstTangenteEn(droite, cercle ou demi-cercle ou ellipse, point)*
- *EstMédiatriceEn(droite, segment, point)*
- *EstDistanceEntre(nombre, point ou LO² ou LF³, point ou LO ou LF)*
- *IntersectentEn(LO ou LF, LO ou LF, point)*
- *AnglesConsécutifs(angle,angle,polygone)*

Dans l'exemple que nous avons utilisé pour illustrer les phases précédentes, cette phase n'ajoutera pas de nouveaux éléments puisque le problème ne contient pas des fonctions et de prédicats ternaires.

² Ligne ouverte comme les droites, les segments, les arcs, etc.

³ Ligne fermée comme le cercle, les polygones, les ellipses, etc.

Prenons l'exemple suivant qui contient un prédicat ternaire : 'la droite d est sécante au cercle c au point A'. A la sortie de ce module nous aurons les hypothèses suivantes :

- *EstCercle(c)*
- *EstDroite(d)*
- *EstPoint(A)*
- *IntersectentEn(d,c,A)* : Un prédicat ternaire que nous pouvons l'avoir grâce à ce module.

5.3.5. Extraction des Conclusions

Nous différencions cinq types de conclusions :

- Les conclusions qui sont sous forme d'une **question ouverte**. Par exemple, les questions qui commencent par les déterminants interrogatifs *quel, quelle, quels* et *quelles* ou qui commencent par le mot **trouver**. Le but de cette conclusion est de trouver soit la nature d'une entité géométrique soit une valeur ou une mesure.
- Les conclusions qui sont sous forme d'une **question fermée**, par exemple les questions qui commencent par *Est-ce que*.
- Les conclusions qui sont sous forme d'une **question à choix multiples**, par exemple les questions qui commencent par *parmi ... lequel(le)...*
- Les conclusions qui demandent de **démontrer** une propriété, par exemple les questions qui commencent par *démontrer, prouver et montrer*.

5.3.5.1. Questions ouvertes

Les conclusions qui sont sous forme d'une question ouverte ont pour but de déterminer la valeur ou la nature d'une entité géométrique présente dans le texte.

- Si nous cherchons à déterminer la valeur numérique d'une entité, nous exprimons la conclusion à l'aide du prédicat *TrouverValeur(nombre)*. Prenons en exemple l'expression "Quelle est la longueur du segment [AB] ?", l'hypothèse sera formulée comme *TrouverValeur(AB)*. Ce prédicat informe le résolveur qu'il doit trouver la mesure **AB** du segment **[AB]**. En général, dans ce type de question nous cherchons : la mesure d'un segment, la mesure d'un arc, le périmètre d'un polygone, l'aire d'un polygone ou la mesure d'un angle. Par conséquent, une question de ce type qui est résolue par le générateur de preuve de QED-Tutrix aura en sortie de ce dernier une conclusion de la forme *EstMesureDe(entité,valeur)*, où *entité* est un item de l'ensemble

{angle, segment, arc, polygone}, et où *valeur* représente la valeur numérique trouvée par le générateur de preuves.

- Si on cherche à déterminer la nature d'une entité, nous exprimons la conclusion par la propriété : *TrouverNature(entité)*. Par exemple, la conclusion de la phrase '*Quelle est la nature du triangle ABC*', sera formulée *TrouverNature(ABC)*. Dans ce cas, le résolveur est informé qu'il doit trouver la nature du triangle **ABC**.

En général, dans ce genre de question, nous cherchons soit la nature d'un triangle, soit la nature d'un quadrilatère. Par conséquent, la réponse à cette question sera :

- *TrouverNature(triangle)* : Les réponses possibles de cette conclusion seront parmi une des sous-classes du triangle suivantes : scalène, isocèle, équilatéral, isoangle, acutangle, équiangle, obtusangle et rectangle. Par conséquent, dans le cas où une question de ce type est donnée en entrée au générateur automatique de preuves de QED-Tutrix, nous trouvons parmi les inférences générées en sortie une conclusion de la forme suivante : *Prédicat(triangle)*, où *Prédicat* est un prédicat de l'ensemble suivant: {*EstIsocèle*, *EstEquilatéral*, *EstIsoangle*, *EstAcuangle*, *EstEquiangle*, *EstObtusangle*, *EstRectangle*}.
- *TrouverNature(quadrilatère)* : Les réponses possibles de cette conclusion seront une des sous-classes du quadrilatère suivantes : parallélogramme, losange, trapèze, carré, cerf-volant et rectangle. Le générateur de preuve de QED-Tutrix infère pour ce type de question une conclusion de la forme *Prédicat(quadrilatère)*, où *Prédicat* est un prédicat de l'ensemble suivant: {*EstLosange*, *EstTrapèze*, *EstCerfVolant*, *EstCarré*, *EstRectangle*}.

5.3.5.2. Questions fermées

Dans ce cas, nous utilisons le prédicat *Conclusion(hypothèse)*. Il permet de préciser au résolveur qu'une *hypothèse* est une conclusion et non pas une donnée pour la résolution de la question. Par exemple, prenons la phrase '*Est-ce que le triangle ABC est isocèle ?*'. Avec le générateur des prédicats binaires, nous trouvons l'hypothèse *EstIsocèle(ABC)*. Nous encapsulons cette hypothèse dans la prédicat *Conclusion(EstIsocèle(ABC))* pour préciser au générateur de preuves qu'il s'agit d'une conclusion que nous devons trouver dans les inférences produites.

Pour les questions de démonstration, nous utilisons le même prédicat *Conclusion(hypothèse)*, afin de préciser au résolveur qu'une hypothèse est une conclusion à démontrer. Par exemple,

l'expression "Démontrer que le triangle ABC est rectangle" se formule ainsi *Conclusion(EstTriangleRectangle(ABC))*.

Les prédicats de conclusions sont générés et classifiés en même temps que les prédicats d'hypothèses et avec le même processus. Par exemple, les prédicats *Conclusion(hypothèse)*, *TrouverValeur(nombre)*, *TrouverNature(polygone)* sont traités comme des prédicats unaires, tel que décrit à la section 5.3.2.

Par exemple, la phrase '*quelle relation y a-t-il entre les aires des triangles AEB et AED ?*' de l'exemple 1 permet d'extraire la conclusion suivante :

TrouverValeur(Rapport(AireDe(AED),AireDe(AED))

5.3.5.3. Questions à choix multiples

Nous procédons de la même façon que nous avons fait dans les questions fermées, mais cette fois avec une propriété qui prend en argument plusieurs hypothèses : *Conclusions(*hypothèse)*. Par exemple, considérons la phrase suivante '*Trouvez la bonne réponse : le triangle ABC est isocèle en A; le triangle ABC est isocèle en B; le triangle ABC est isocèle en C.*'. Cette expression est formulée comme suit : *Conclusions(EstIsocèleEn(ABC,B), EstIsocèleEn(ABC,B), EstIsocèleEn(ABC,C))*. Ce prédicat informe le résolveur qu'une de ces trois conclusions est bonne, autrement dit, nous devons trouver dans les inférences générées par le générateur de preuves une de ces conclusions. Actuellement, dans le projet QED-Tutrix, nous ne traitons pas encore ce genre de question.

5.3.6. Extraction des hypothèses implicites et optimisation de l'ensemble des hypothèses extraites

Dans les exemples 2 et 4 du corpus MSD, il y a des hypothèses implicites que ni GEOS ni GEOSFR ne sont capables d'extraire. Dans l'exemple 2, *Démontrer que le quadrilatère ABCD ayant 3 angles droits est un rectangle*, les trois angles font référence aux trois angles parmi les quatre angles {ABC, BCD, CDA, DAB} du quadrilatère ABCD. De même pour l'exemple 4, *Soit ABC un triangle quelconque, on appelle g la bissectrice issue du sommet B et f la bissectrice issue du sommet C...*, f et g font référence aux bissectrices des angles {ABC, BCA}. Pour résoudre ce problème, nous avons créé un ensemble de règles (voir le tableau 18 qui présente quelques exemples de ces règles) permettant de trouver les hypothèses implicites qui manquent ou pour créer de nouvelles entités qui ne sont pas présentes directement dans le texte

mais nécessaires au générateur de preuve pour la résolution. Cet ensemble de règles est utilisé par l'algorithme **ERHI** (voir ci-bas). Cet ensemble de règles est ajouté à la sortie des générateurs des prédicats unaires, binaires et ternaires qui utilisent l'apprentissage machine pour extraire les hypothèses explicites des énoncés. Cet ensemble de règles est sujet à modifications, c'est-à-dire que nous pouvons ajouter ou enlever des règles selon les besoins du générateur de preuves. L'avantage de cette approche est qu'il permet à notre extracteur d'être aisément adaptable à différents générateurs de preuves. Prenons l'exemple suivant:

Soient (IS) et (CN) des droites parallèles. Y est un point de du segment [IS] et R est un point du segment [CN] tels que l'angle IYR mesure 72 degrés. La bissectrice de l'angle RYS coupe [CN] en L. On considère les points T et A tels que T appartient au segment [YS], l'angle ATL mesure 84 degrés et l'angle ALR est droit. Détermine la mesure de l'angle TAL.

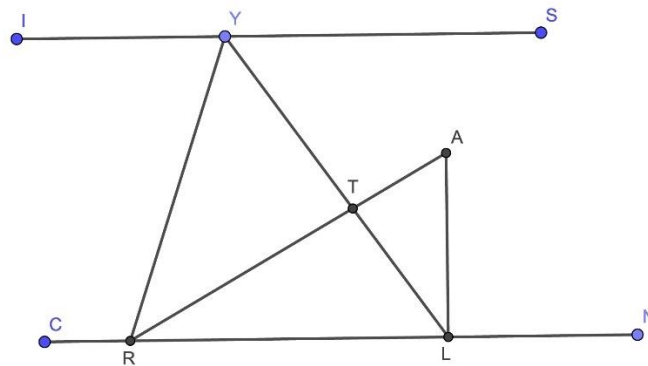


Figure 15 : Figure illustrative de l'exemple.

Les hypothèses et les conclusions extraites de cet exemple sont les suivantes :

- **Hypothèses :**
 - (IS) est parallèle à (CN)
 - Y sur [IS]
 - R sur [CN]
 - L sur [CN]
 - T sur [YS]
 - La bissectrice de l'angle RYS coupe [CN] en L
 - ALR est un angle droit
 - L'angle IYR mesure 72 degrés
 - L'angle ATL mesure 84 degrés

Mesure $ALR=90$ degrés

(YL) est la bissectrice de RYS

L'angle RYL

L'angle LYS

YL est une sécante de (IS) et (CN)

YR est une sécante de (IS) et (CN)

Triangle ATL

- **Conclusions :**

Mesure de l'angle TAL

Cet exemple nous révèle plusieurs difficultés que l'extracteur de connaissances doit traiter :

- Les hypothèses implicites et explicites : Certaines hypothèses ne sont pas dites explicitement dans le texte, comme l'hypothèse 'L sur [CN]'. En fait, les hypothèses 'L sur [CN]' et '(YL) est la bissectrice de RYS' se déduisent de l'hypothèse 'La bissectrice de l'angle RYS coupe [CN] en L'. Un autre exemple d'hypothèses implicites, l'hypothèse Triangle ATL n'est pas mentionnée dans le texte. Nous déduisons son existence du fait que les trois points A, L et T sont trois points distincts reliés par les droites distinctes AL, LT et AT. Or, le triangle ATL doit être présent dans l'entrée du générateur de preuve pour trouver une des solutions de ce problème qui utilise ce triangle. Voir dans le tableau 17 une démonstration de l'exemple ci-dessus qui utilise le triangle ATL qui est une hypothèses implicites. Il faut prévoir que l'extracteur couvre toutes les hypothèses implicites de ce genre. Afin de détecter ce genre d'hypothèses, nous avons utilisé l'ensemble de questions du corpus MSD résolues par des didacticiens, et avons ajouté les règles nécessaires pour extraire les hypothèses implicites qui ont été identifiées par les didacticiens.

Tableau 17 : une démonstration du second exemple.

Affirmation	Justification
Mesure(SYR) = 108	Les angles SYR et IYR sont supplémentaires
Mesure(RYL) = 54	YL est bissectrice de RYS
Mesure(IYL) = 126	IYR et RYL sont adjacents

Mesure(YLN) = 126	Les angles IYL et YLN sont alternes-internes
Mesure(YLA) = 36	Les angles YLA et ALN sont adjacent, Mesure(ALN) = 90 et Mesure(YLN) = 126
Mesure(TAL) = 60	La somme des angles de triangle ATL est 180 degrés.

- L'adaptabilité de l'extracteur à QED-Tutrix : La sortie de l'extracteur doit être adaptable à la représentation des connaissances dans QED-Tutrix. Dans cet exemple, l'hypothèse '*l'angle ALR est droit*' doit être traduite en 'Mesure ALR=90 degrés' afin qu'elle soit adéquate à la forme d'entrée de QED-Tutrix. Cette adaptabilité sera discuté dans la section 5.3.7.

Soit I l'ensemble des règles pour extraire les hypothèses implicites. Une règle de I est une fonction $règle_i(h_1..h_m) \rightarrow H_1..Hp$ qui prend en argument un ensemble d'hypothèses $h_1..h_m$ et produit les hypothèses implicites $H_1..Hp$, où $i = 1, 2..n$ sont les indices des règles dans le tableau 18 et n est le nombre total de ces règles. L'algorithme ERHI ci-dessous décrit le processus permettant d'inférer les hypothèses implicites.

Algorithme ERHI : Extraction des hypothèses implicites en utilisant l'ensemble de règles I

Entrée : l'ensemble des hypothèses explicites HE , l'ensemble des règles $I = \{règle_i(h_1..h_m) \rightarrow H_1..H_p \mid i = 1, 2..n\}$.

Sortie : l'ensemble des hypothèses explicites et implicites HEI

Initialiser l'ensemble des hypothèses explicites et implicites HEI à HE ;

Initialiser l'ensemble EP à l'ensemble vide qui contiendra les couples $(règle_i, (h_1..h_m))$ que nous avons déjà utilisés ;

Tant qu'il existe une $règle_i \in I$ et $(h_1..h_m) \subset HEI$ tel que $(h_1..h_m)$ correspondent aux arguments de $règle_i$ et le couple $(règle_i, (h_1..h_m))$ n'est pas dans l'ensemble EP **Faire:**

Ajouter les hypothèses $H_1..H_p$ à HEI obtenues par l'application de $règle_i$ aux hypothèses $(h_1..h_m)$;

Ajouter le couple $(règle_i, (h_1..h_m))$ à l'ensemble EP ;

Fin

Retourner HEI

Tableau 18 : Une partie de l'ensemble des règles pour l'extraction des hypothèses implicites (ERHI)

Numéro et nom de la règle	Hypothèses de départ $(h_1..h_m)$	Hypothèses implicites $H_1..H_p$
1. Deux point X_1 et X_2	EstPoint(X_1) EstPoint(X_2)	EstDroite(X_1X_2) EstSegment(X_1X_2)
2. Trois points X_1, X_2 et X_3	EstPoint(X_1) EstPoint(X_2) EstPoint(X_3)	EstTriangle($X_1X_2X_3$)
3. Quatre points X_1, X_2, X_3 et X_4	EstPoint(X_1) EstPoint(X_2) EstPoint(X_3) EstPoint(X_4)	EstQuadrilatère($X_1X_2X_3X_4$)
4. n point $X_1, X_2..X_n$	EstPoint(X_1) .. EstPoint(X_n)	
5. Polygone $X_1X_2..X_n$	EstPolygone($X_1X_2..X_n$)	EstSommetDe($X_i, X_1X_2..X_n$) EstAngleDe($X_{i-1}X_iX_{i+1}, X_1X_2..X_n$) EstCôtéDe($X_iX_{i+1}, X_1X_2..X_n$) EstDiagonalDe($X_iX_j, X_1X_2..X_n$) SommetsAdjacents(X_i, X_{i+1}) CôtésAdjacents($X_iX_{i+1}, X_{i+1}X_{i+2}$)

		AnglesAdjacents($X_{i-1}X_iX_{i+1}$, $X_iX_{i+1}X_{i+2}$)
6. Polygone régulier $X_1X_2..X_n$	EstPolygone($X_1X_2..X_n$)	EstPolygone($X_1X_2..X_n$) DeMêmeMesure(MesureDe($X_{i-1}X_iX_{i+1}$), Mesure($X_iX_{i+1}X_{i+2}$)) DeMêmeMesure(MesureDe($X_{i-1}X_i$), MesureDe(X_iX_{i+1}))
7. Polygone d'ordre pair $X_1X_2..X_n$	EstPolygonePair($X_1X_2..X_n$)	EstPolygone($X_1X_2..X_n$) EstPair(n) SommetsOpposés($X_i, X_{i+\frac{n}{2}}$) Opposés($X_iX_{i+1}, X_{i+\frac{n}{2}}X_{i+\frac{n}{2}+1}$) SontOpposés($X_{i-1}X_iX_{i+1}$, $X_{i+\frac{n}{2}-1}X_{i+\frac{n}{2}}X_{i+\frac{n}{2}+1}$)
8. Polygone d'ordre impair $X_1X_2..X_n$	EstPolygoneimpair($X_1X_2..X_n$)	EstImpair(n) SommetsOpposés(X_i , $X_{i+\frac{n-1}{2}}X_{i+1+\frac{n-1}{2}}$)
9. Cerf- volant $X_1X_2X_3X_4$	EstCerfVolant($X_1X_2X_3X_4$)	EstAxeDeSymétrie(X_1X_3) EstAxeDeSymétrie(X_2X_4)
10. Trapèze $X_1X_2X_3X_4$	EstTrapèze($X_1X_2X_3X_4$)	Parallèle(X_1X_2, X_3X_4) ou Parallèle(X_2X_3, X_1X_4)
11. parallélogramme $X_1X_2X_3X_4$	EstParllélogramme($X_1X_2X_3X_4$)	Parallèle(X_1X_2, X_3X_4) Parallèle(X_2X_3, X_1X_4) DeMêmeMesure(MesureDe(X_1X_2), MesureDe(X_3X_4)) DeMêmeMesure(MesureDe(X_2X_3), MesureDe(X_1X_4)) DeMêmeMesure(MesureDe($X_1X_2X_3$), MesureDe($X_3X_4X_1$)) DeMêmeMesure(MesureDe($X_2X_3X_4$), MesureDe($X_4X_1X_2$))
12. Rectangle $X_1X_2X_3X_4$	EstQuadRectangle($X_1X_2X_3X_4$)	DeMêmeMesure(MesureDe($X_1X_2X_3$), MesureDe($X_3X_4X_1$)) DeMêmeMesure(MesureDe($X_3X_4X_1$), MesureDe($X_2X_3X_4$)) DeMêmeMesure(MesureDe($X_2X_3X_4$), MesureDe($X_4X_1X_2$))

		DeMêmeMesure(MesureDe($X_4X_1X_2$),90)
13. Losange $X_1X_2X_3X_4$	EstLosange($X_1X_2X_3X_4$)	DeMêmeMesure(MesureDe(X_1X_2), MesureDe(X_2X_3)) DeMêmeMesure(MesureDe(X_2X_3), MesureDe(X_3X_4)) DeMêmeMesure(MesureDe(X_3X_4), MesureDe(X_1X_4))
14. Carré $X_1X_2X_3X_4$	EstCarré($X_1X_2X_3X_4$)	DeMêmeMesure(MesureDe(X_1X_2), MesureDe(X_2X_3)) DeMêmeMesure(MesureDe(X_2X_3), MesureDe(X_3X_4)) DeMêmeMesure(MesureDe(X_3X_4), MesureDe(X_1X_4)) DeMêmeMesure(MesureDe($X_1X_2X_3$), MesureDe($X_3X_4X_1$)) DeMêmeMesure(MesureDe($X_3X_4X_1$), MesureDe($X_2X_3X_4$)) DeMêmeMesure(MesureDe($X_2X_3X_4$), MesureDe($X_4X_1X_2$)) DeMêmeMesure(MesureDe($X_4X_1X_2$),90)
15. Triangle isocèle $X_1X_2X_3$ en X_i	EstIsocèleEn($X_1X_2X_3, X_i$)	DeMêmeMesure(MesureDe(X_iX_{i+1}), MesureDe(X_iX_{i-1})) DeMêmeMesure(MesureDe($X_iX_{i+1}X_{i-1}$), MesureDe($X_iX_{i-1}X_{i+1}$))
16. Triangle équilateral $X_1X_2X_3$	EstÉquilatéral($X_1X_2X_3$)	DeMêmeMesure(MesureDe(X_1X_2), MesureDe(X_2X_3)) DeMêmeMesure(MesureDe(X_2X_3), MesureDe(X_3X_1)) DeMêmeMesure(MesureDe($X_1X_2X_3$), MesureDe($X_2X_3X_1$)) DeMêmeMesure(MesureDe($X_2X_3X_1$), MesureDe($X_3X_1X_2$)) DeMêmeMesure(60, MesureDe($X_3X_1X_2$))
17. Triangle rectangle $X_1X_2X_3$ en X_i	EstRéctangleEn($X_1X_2X_3, X_i$)	DeMêmeMesure(90, MesureDe($X_{i-1}X_iX_{i+1}$)) EstDroit($X_{i-1}X_iX_{i+1}$)
18. Bissectrice X_2X_4 de $X_1X_2X_3$	EstBissectriceDe($X_2X_4, X_1X_2X_3$)	DeMêmeMesure(MesureDe($X_1X_2X_4$), MesureDe($X_4X_2X_3$))

19. Y est la bissectrice issue de X_i d'un polygone $X_1X_2 \dots X_n$	$\text{BissectriceIssueDe}(Y, X_iX_1X_2 \dots X_n)$	$\text{EstBissectriceDe}(Y, X_{i-1}X_iX_{i+1})$ $\text{Appartient}\hat{A}(X_i, Y)$
20. LO est une ligne ouverte sécante à LF une ligne fermée	$\text{EstSécanteDe}(LO, LF)$	$\text{Appartient}\hat{A}(X_1, LO)$ $\text{Appartient}\hat{A}(X_2, LO)$ $\text{Appartient}\hat{A}(X_1, LF)$ $\text{Appartient}\hat{A}(X_2, LF)$
21. $LO1$ est une ligne ouverte sécante à $LO2$ une ligne Ouverte	$\text{EstSécanteDe}(LO1, LO2)$	$\text{Appartient}\hat{A}(X, LO1)$ $\text{Appartient}\hat{A}(X, LO2)$
22. LO est une ligne ouverte tangente à LF une ligne fermée	$\text{EstTangenteDe}(LO, LF)$	$\text{Appartient}\hat{A}(X, LO)$ $\text{Appartient}\hat{A}(X, LF)$
23. X_1X_2 est sécante à X_3X_4 en Y	$\text{EstSécanteDeEn}(X_1X_2, X_3X_4, Y)$	$\text{DeMêmeMesure}(\text{MesureDe}(X_1YX_3), \text{MesureDe}(X_2YX_4))$ $\text{DeMêmeMesure}(\text{MesureDe}(X_1YX_4), \text{MesureDe}(X_2YX_3))$
24. X_1, X_2 et X_3 trois points alignés	$\text{SontAlignés}(X_1, X_2, X_3)$	$\text{Appartient}\hat{A}(X_1, X_2X_3)$ $\text{Appartient}\hat{A}(X_2, X_1X_3)$ $\text{Appartient}\hat{A}(X_3, X_1X_2)$
25. X_1X_2 est médiatrice de X_2X_4	$\text{EstMédiatriceDe}(X_1X_2, X_3X_4)$	$\text{EstMédiatriceDeEn}(X_1X_2, X_3X_4, Y)$
26. X_1X_2 est médiatrice de X_2X_4 en Y	$\text{EstMédiatriceDeEn}(X_1X_2, X_3X_4, Y)$	$\text{EstSécanteDe}(X_1X_2, X_3X_4)$ $\text{DeMêmeMesure}(\text{MesureDe}(X_1YX_3), \text{MesureDe}(X_2YX_4))$ $\text{DeMêmeMesure}(\text{MesureDe}(X_2YX_4), \text{MesureDe}(X_1YX_4))$ $\text{DeMêmeMesure}(\text{MesureDe}(X_1YX_4), \text{MesureDe}(X_2YX_3))$ $\text{DeMêmeMesure}(\text{MesureDe}(X_2YX_3), 90)$
27. m est la médiane issue de X_i dans le triangle $X_1X_2X_3$	$\text{MédianeIssueDe}(m, X_i, X_1X_2X_3)$	$\text{Appartient}\hat{A}(Y, X_{i-1}X_{i+1})$ $\text{DeMêmeMesure}(\text{MesureDe}(X_{i-1}Y), \text{MesureDe}(YX_{i+1}))$
28. Le cercle C	$\text{EstCercle}(C)$	$\text{EstCentreDe}(O, C)$ $\text{Appartient}\hat{A}(X, C)$

		$\text{EstRayonDe}(OX, C)$ $\text{Appartient}\hat{A}(Y, OX)$ $\text{Appartient}\hat{A}(Y, C)$ $\text{Distinct}(X, Y)$ $\text{EstDiamètreDe}(XY, C)$
29. Le cercle C de diamètre XY de centre O	$\text{EstCercle}(C)$ $\text{EstCentreDe}(O, C)$ $\text{EstDiamètreDe}(XY, C)$	$\text{Appartient}\hat{A}(O, XY)$
30. h est la hauteur issue de X_i dans le triangle $X_1X_2X_3$	$\text{HauteurIssueDe}(h, X_i, X_1X_2X_3)$	$\text{Appartient}\hat{A}(Y, X_{i-1}X_{i+1})$ $\text{DeMêmeMesure}(\text{MesureDe}(X_{i-1}YX_i), \text{MesureDe}(X_iYX_{i+1}))$
31. h est la hauteur issue de X_i dans le parallélogramme $X_1X_2X_3X_4$	$\text{HauteurIssueDe}(h, X_i, X_1X_2X_3X_4)$	$\text{Appartient}\hat{A}(Y, X_{i-1}X_{i+2})$ $\text{DeMêmeMesure}(\text{MesureDe}(X_{i-1}YX_i), \text{MesureDe}(X_iYX_{i+2}))$
32. m angles droits d'un polygone $X_1X_2 \dots X_n$	$\text{mAnglesDuPolygone}(m, \text{angles } X_1X_2 \dots X_n)$ $\text{EstDroits}(\text{angles})$ Remarque : de même si les angles sont par exemple aigus, plein etc	$\text{EstDroite}(X_1X_2X_3)$ $\dots$ $\text{EstDroite}(X_mX_{m+1}X_{m+2})$
	...	...

Pour résumer le tableau 18 ci-dessus, l'extracteur des hypothèses implicites a plusieurs buts :

- Résoudre les problèmes issus des coréférences présents dans le texte. Citons la phrase 'Soit ABC un triangle quelconque, on appelle g la bissectrice issue du sommet B et f la bissectrice issue du sommet C .' de l'exemple 4, où nous pouvons utiliser la règle 19 pour déterminer les angles des bissectrices.
- Dédire les mesures des angles reliés aux médiatrices, les bissectrices, les hauteurs, etc.
- Dédire des propriétés sur les polygones, comme le carré, le triangle isocèle, losange, etc. Par exemple, les angles du carré sont tous égaux à 90 degrés.
- Explorer de nouvelles entités géométriques implicites qui ne sont pas mentionnées dans le texte, mais qui résultent de la construction géométrique, et que nous pouvons normalement voir quand nous faisons la construction figurale des données du texte. Par exemple, à partir d'un ensemble de points, nous pouvons inférer des triangles et des polygones qui ne figurent pas dans le texte. Ce dernier type d'hypothèse implicite est

ajouté afin de fournir au résolveur toutes les données géométriques possibles pour la résolution.

Revenons à notre exemple où nous avons à l'entrée de ce module les hypothèses suivantes :

- *EstPoint(E)*
- *EstDiagonale(AC)*
- *EstParallélogramme(ABCD)*
- *EstTriangle(AEB)*
- *EstDiagonaleDe(AC,ABCD)*
- *EstTriangle(AED)*
- *TrouverValeur(Rapport(AireDe(AED),AireDe(AED)))*

A la sortie de ce module, s'ajoutent plus de 1000 hypothèses implicites (dans le cas où toutes les règles du tableau 18 sont appliquées). En voici quelques-unes :

- *EstSegment(AE)*
- *EstSegment(AC)*
- *EstSegment(AD)*
- *EstTriangle(ABC)*
- *EstTriangle(ABC)*
- *EstTriangle(ABE)*
- *EstQuadrilaère(ADEB)*
- ...

Conjonction de coordination :

Comme nous l'avons souligné précédemment, la façon dont GEOS traite les conjonctions de coordination est incomplète. Rappelons que nous avons ce problème dans des questions qui ont des expressions de type 'A et B sont, respectivement, les milieux des segments CD et EG' ou 'les mesures des segments AB, CD, EF et GH sont égales, respectivement, à 5, 6, 7 et 10'. Dans le premier cas, les hypothèses extraites sont *EstMilieuDe(A,CD)*, *CC(A,B)* et *CC(CD,EG)*. Rappelons que l'algorithme *GREEDY_MAXIMAZATION* [4] permettant de choisir l'ensemble optimal final des prédicats en couvrant un maximum de concepts présents dans l'énoncé tout en évitant la répétition, a un désavantage dans ce type de questions. Puisque

la répétition est pénalisée par *GREEDY_MAXIMAZATION*, nous ne pouvons pas avoir dans la sortie des prédicats qui se répètent. Ce qui explique l'absence du prédicat *EstMilieuDe(B,EG)* dans l'exemple ci-dessus. Pour remédier à ce problème, GEOS utilise les prédicats *CC*. Dans le cas ci-dessus, à la sortie de GEOS *EstMilieuDe(A,CD)*, *CC(A,B)* et *CC(CD,EG)* seront transformés en *EstMilieuDe({A,B},{CD,EG})* (Les variables entre accolades ne sont pas ordonnées, c'est-à-dire $\{A,B\}$ est équivalent à $\{B,A\}$). Toutefois, le problème n'est pas complètement résolu parce que nous ignorons avec ce prédicat qui est le milieu de quel segment. Pour résoudre ce problème, nous utilisons dans GEOS++ l'ordre d'apparition des variables dans le texte. Ainsi, dans l'expression 'A et B sont, respectivement, les milieux des segments CD et EG', A apparaît avant B, de même pour CD qui apparaît avant EG. Cet ordre est utilisé pour associer A à CD et B à EG. Ce qui permet d'avoir en sortie *EstMilieuDe(A,CD)* et *EstMilieuDe(B,EG)* qui est plus précis que *EstMilieuDe({A,B},{CD,EG})*. Pour ce faire, nous utilisons un ensemble de règles pour la conjonction de coordination. Voir quelques exemples de ces règles dans le tableau 19. En total, nous avons 26 règles pour traiter les conjonctions de coordinations.

Tableau 19 : Une partie de l'ensemble des règles pour l'extraction des hypothèses implicites (ERHI) concernant les conjonctions de coordination

Le nom de la règle	Hypothèses de départ ($h_1 \dots h_m$)	Hypothèses implicites $H_1 \dots H_p$
61. A et F sont, respectivement, les milieux de BC et ED	<i>EstMilieuDe(A,BC)</i> <i>CC(A,F)</i> <i>CC(BC,ED)</i> $\text{indice}^4(A) < \text{indice}(F)$ $\text{indice}(BC) < \text{indice}(ED)$	<i>EstMilieuDe(A,BC)</i> <i>EstMilieuDe(F,ED)</i>
62. Les mesures de AB, CD et EF sont égales, respectivement, à V1, V2 et V3	<i>EstMesureDe(V1,AB)</i> <i>CC(V1,V2)</i> <i>CC(V1,V3)</i> <i>CC(AB, CD)</i> <i>CC(AB, EF)</i> $\text{indice}(V1) < \text{indice}(V2) < \text{indice}(V3)$ $\text{indice}(AB) < \text{indice}(BC) < \text{indice}(EF)$	<i>EstMesureDe(V1,AB)</i> <i>EstMesureDe(V2,CD)</i> <i>EstMesureDe(V3,EF)</i>
...		

⁴ indice est l'indice de l'apparition de la variable dans le texte (voir la section 4.2, où nous donnons des exemples avec des indices)

5.3.7. Adaptation de l'extracteur au générateur de preuve QED-Tutrix

Le dernier module des hypothèses implicites permet d'adapter la sortie de l'extracteur afin qu'il puisse être traité par le générateur de preuves de QED-Tutrix. Une fois qu'on a les hypothèses et les conclusions en sortie du dernier module des hypothèses implicites représentées par Ω_+ , il faut les traduire dans la représentation logique de QED-Tutrix. Voici quelques exemples : `EstPoint(A)`, `EstCentreDe(O,C)` et `Parallèle(L1,L2)` qui sont écrit avec le formalisme Ω_+ , ils seront convertis, respectivement, dans la représentation de QED-Tutrix, en `point(A)`, `circleCenter(O,C)`, `para(line(L1),line(L2))`. Cependant, la conversion n'est pas directe dans quelques cas que nous présentons dans la suite. En total, nous utilisons 60 règles de conversions entre les deux représentations, y inclus les cas spéciaux.

Cas spéciaux de la conversion :

- **Droite** : une droite dans la représentation logique de QED-Tutrix est représentée par une liste de points sans ordre. Par exemple, une droite qui contient les points A, B, C et D est représenté par `Line([A,B,C,D])`. Il faut donc connaître tous les points qui appartiennent à une droite quelconque avant de la convertir. Ce qui demande de traiter les intersections des lignes, les tangentes, les prédicats *AppartientA(point,droit)*, *EstMilieuDe(point,droite)*, *EstBissectricesDe*, *EstMédiatriceDe* et tous ce qui impliquent l'intersection d'une droite avec d'autres entités géométriques. Une fois les

points d'une droite connus, nous devons substituer toutes ces droites avec les listes des points.

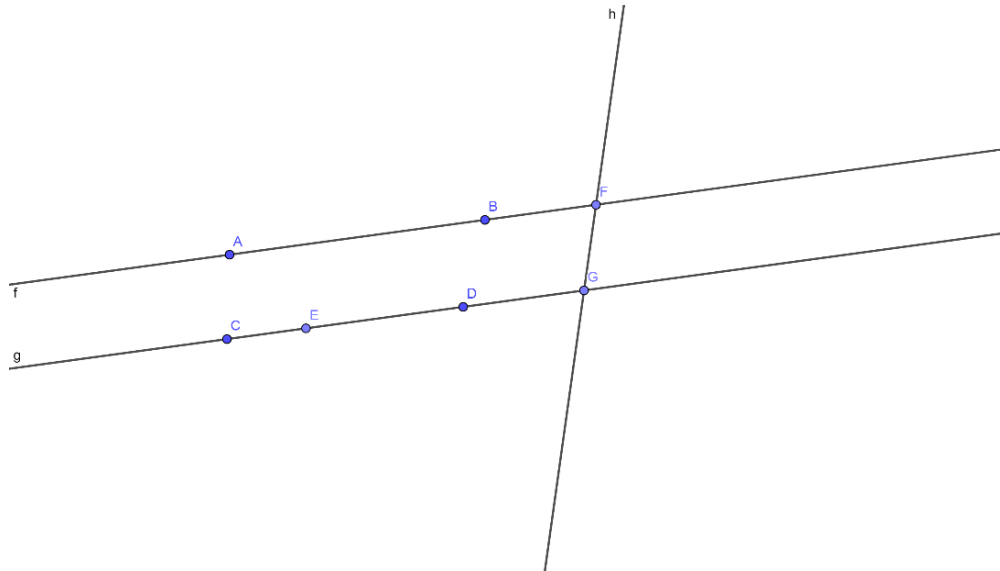


Figure 16 : Un exemple de représentation figurale de deux droites parallèles qui intersectent une troisième droite

Prenons par exemple la situation illustrée à la figure 16, qui est représentée par les hypothèses suivantes : $Parallèle(AB,CD)$, $AppartientA(E,CD)$, $IntersectentEn(CD,FG,G)$ et $IntersectentEn(AB,FG,F)$. Celles-ci seront converties comme suit : $Para([A,B,F],[C,D,E,G])$, $line([A,B,F])$, $line([C,D,E,G])$ et $line([H,F])$. Pour ce faire, nous avons ajouté dans les règles pour l'extraction des hypothèses implicites des nouvelles règles qui permettent de trouver tous les points d'une droite donnée. Une fois cette opération est faite, les droites sont remplacées dans les hypothèses trouvées par des listes de points.

- **Angle** : Un angle dans QED-Tutrix est représenté comme suit : $angle(liste_de_points, point, liste_de_points)$, où chaque liste de points représente un des deux segments de droites définissant l'angle. La position des liste de points (en premier ou troisième argument du prédicat) est importante. Dans la figure 17, nous donnons quelques exemples d'angles avec la représentation de QED-Tutrix : l'angle vert est représenté par $angle([C,E],A,[D,B])$, l'angle rouge est $angle([D,B],A,[F])$ et l'angle bleu est $angle([F],A,[G])$. Dans GEOSFR++, les angles sont représentés différemment, parce qu'ils sont extraits directement du texte. Par exemple, l'angle vert peut être représenté par l'un des prédicats suivants : $Angle(CAD)$, $Angle(CAB)$, $Angle(EAD)$ ou $Angle(EAB)$. Ces quatre angles sont implicitement les mêmes angles. Par conséquent,

dans ce cas, la conversion de la représentation en Ω_+ à la représentation de QED-Tutrix est infaisable parce qu'elle nécessite des informations topologiques (comme, l'orientation des angles, l'ordre des points sur une droite) qui sont présentes uniquement dans la figure et non pas dans le texte. Ce qui nécessite une intervention humaine pour faire des corrections sur les angles après l'extraction pour ce type de question. Par conséquent, nous sommes incapables de traiter les angles pour l'instant. Nous présentons dans le chapitre suivant de l'évaluation les conséquences de cela.

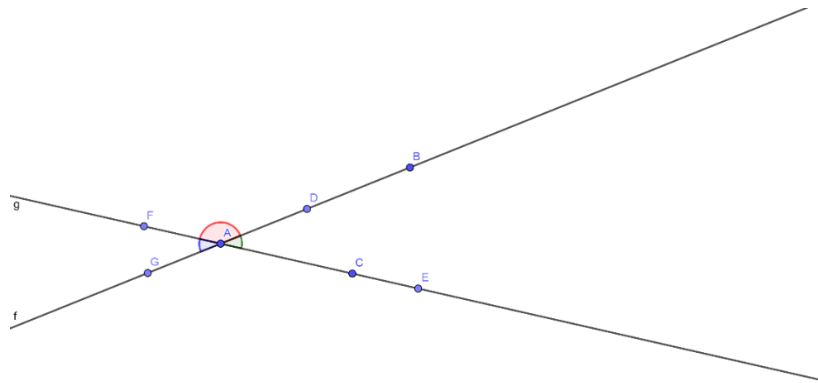
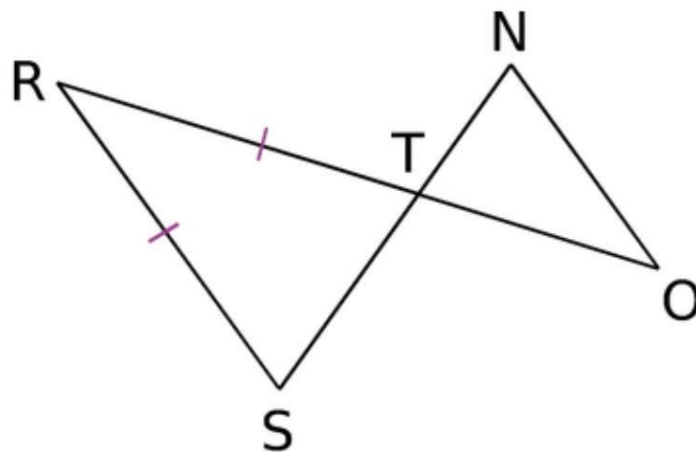


Figure 17 : Exemple de représentation figurale d'angles : $angle([C,E],A,[D,B])$ en vert, $angle([D,B],A,[F])$ en rouge et $angle([F],A,[G])$ en bleu.

Prenons en exemple la question suivante :

Les droites RO et SN sont sécantes en T. Le triangle RST est isocèle en R. Les droites RS et NO sont parallèles. Montre que le triangle TNO est isocèle en O.



Pour que le générateur de preuve soit capable de trouver les solutions de cette question, il est nécessaire de corriger les angles $Angle([r],s,[t])$ et $Angle([t],n,[o])$ en $Angle([r],s,[t,n])$ et $Angle([s,t],n,[o])$.

Générateur de preuve de QED-Tutrix incomplet :

Certains concepts géométriques ne sont pas encore implémentés dans le générateur de preuve de QED-Tutrix, ce qui veut dire qu'une question qui contient ces concepts ne sera pas résolue par le générateur de preuve.

Dans le tableau 20, nous présentons les statistiques sur les concepts géométriques de QED-Tutrix et Ω_+ .

Tableau 20 : Statistique sur les concepts géométriques de QED-Tutrix et Ω_+ .

	Nombres de concepts
Les concepts présents dans Ω_+	94
Les concepts présents dans QED-Tutrix	44
Les concepts présents dans QED-Tutrix mais pas dans Ω_+	0
Les concepts présents dans Ω_+ mais pas dans QED-Tutrix	50

Dans le tableau 21, nous donnons la liste des concepts qui sont présents dans Ω_+ mais pas dans QED-Tutrix.

Tableau 21 : Les concepts géométriques présents dans Ω_+ mais pas dans QED-Tutrix.

	Les concepts géométriques
Les concepts géométriques présents dans Ω_+ mais pas dans QED-Tutrix	pentadécagone pentagone hendécagone heptagone dodécagone décagone hexagone octogone cerfVolant trapèze scalène isoangle acutangle équiangle obtusangle transformation périmètre circonférence rapport homothétie agrandissement réflexion rotation horaire antihoraire translation surface aigu obtus foyer orthocentre centreDeGravité cathète axe tangente sécante axe axedeRéflexion axedeSymétrie axevertical axehorizontal axefocal

	petitAxe grandAxe demiCercle ellipse étoile polygonesOrdrePair polygonesOrdreImpair aire
--	---

CHAPITRE 6 ÉVALUATION

Dans cette section, nous présentons les résultats de l'évaluation de l'extracteur GEOS++. Nous présentons les performances en précision et rappel de l'extracteur GEOSFR++ sur la validation croisée et sur les données de test. Ensuite, nous présentons les résultats de la résolution des problèmes de l'ensemble de test en utilisant le générateur de preuve QED-Tutrix.

6.1. La validation croisée (rappel et précision)

Dans tous les résultats de cette section, nous avons utilisé la validation croisée tel qu'expliqué plus tôt sur le corpus MSD constitué de 140 questions :

- 60 questions choisies par des didacticiens et inspirées de ce que nous trouvons dans le matériel didactique des enseignants.
- 50 questions SAT [4].
- 30 questions tirées des manuels scolaires.

Tableau 22 : comparaison des résultats de GEOSFR et GEOSFR++

Résultats de l'extraction	Précision	Rappel	F-mesure
GEOSFR (SAT + questions des didacticiens + questions des manuels)	40%	45%	42%
GEOSFR (Questions des didacticiens + Questions des manuels)	30%	28%	28%
GEOSFR++ (SAT + questions des didacticiens + questions des manuels)	85%	90%	87%

GEOSFR++ (Questions des didacticiens + Questions des manuels)	87%	94%	90%
---	-----	-----	-----

Dans le tableau 22, nous remarquons que GEOSFR++ donne des bons résultats par rapport aux résultats de GEOSFR trouvés dans la section 4.2.1. Les résultats s'améliorent dans le cas où nous prenons en considération ou pas les questions SAT dans le calcul. Nous remarquons que les résultats sont améliorés considérablement pour les questions choisies par les didacticiens et les questions des manuels scolaires. Nous passons de 28% à 90% pour la F-mesure.

Les erreurs d'extraction résultent dans les deux cas pour GEOSFR++ en une grande partie de la présence des notions dans l'ensemble de validation qui ne sont pas présentes dans l'ensemble d'entraînement. En fait, dans le générateur des étiquettes, un dictionnaire d'étiquetage qui relie les mots du texte avec les concepts correspondants de la logique Ω_+ , est constitué à partir de l'ensemble d'entraînement. Dans le cas où un concept n'est pas vu dans l'entraînement, il sera ignoré dans la validation. Nous pouvons améliorer ces résultats en ajoutant plus de questions qui sont plus variées et qui couvrent le maximum de concepts.

La conjonction de coordination :

Comme nous avons décrit ci-dessus, nous avons amélioré le processus de la conjonction de coordination utilisé par GEOS en tenant compte de l'ordre d'apparition des concepts dans le texte. Le tableau 23 présente les résultats de cette amélioration.

Tableau 23 : Comparaison entre GEOSFR++ sans et avec le processus de la finalisation par conjonction de coordination amélioré

Résultats de l'extraction (Validation croisée sur les 140 questions du MSD)	Précision	Rappel	F-mesure	Le nombre moyen d'hypothèses et de conclusions ⁵	Le nombre moyen d'hypothèses et de conclusions extraites	Le nombre moyen d'hypothèses et de conclusions correctement extraites
---	-----------	--------	----------	---	---	---

⁵ Le nombre moyen d'hypothèses et de conclusions dans 28 questions de validation qui constitue un cinquième de 140 questions de MSD (validation croisée à cinq plis). Le nombre moyen d'hypothèses et de conclusions par question est 18.

GEOSFR++ avant l'amélioration des conjonctions de coordination	60%	70%	64%	504	587	353
GEOSFR++ après l'amélioration des conjonctions de coordination	85%	90%	87%	504	534	454

Nous remarquons que le traitement de la conjonction de coordination de GEOSFR++ a permis d'améliorer les résultats.

La correction des angles :

Comme nous avons expliqué précédemment, le générateur de preuve nécessite une forme spécifique pour les angles incluant des informations figurales (orientation des angles et l'ordre des points sur une droite) qui ne sont pas en général présentes dans l'énoncé textuel. Actuellement, l'extracteur est incapable de traiter correctement les angles, ce qui nécessite une correction manuelle des angles à la sortie.

Tableau 24 : Comparaison des résultats de GEOSFR++ avant et après la correction des angles

Résultats de l'extraction (Validation croisée sur les 140 questions du MSD)	Précision	Rappel	F-mesure	Le nombre moyen d'hypothèses et de conclusions	Le nombre moyen d'hypothèses et de conclusions extraites	Le nombre moyen d'hypothèses et de conclusions correctement extraites
GEOSFR++ avant la correction des angles	73%	85%	78%	504	433	368
GEOSFR++ après la correction des angles	85%	90%	87%	504	534	454

6.2. Résultats sur l'ensemble de test

Pour le test final, nous entraînons notre modèle sur les 140 questions du corpus MSD, et nous testons sur les 41 questions de l'ensemble de test tel que décrit dans la sections 3.3.3.

6.2.1. Rappel et précision

Dans le tableau 25, nous comparons les résultats entres GEOSFR++ sans traitement des conjonctions des coordination, sans corrections des angles et GEOSFR++ final.

Tableau 25 : Les résultats de GEOSFR++ sur l'ensemble de test

Résultats de l'extraction	Précision	Rappel	F-mesure
GEOSFR++ sans traitement des conjonctions de coordination	59%	65%	61%
GEOSFR++ sans correction des angles	69%	72%	70%
GEOSFR++	71%	76%	73%

En analysant les erreurs d'extraction, ces derniers résultent principalement de :

- La présence des relations géométriques dans les questions de test qui n'ont pas leurs équivalents dans le formalisme Ω_+ . Par exemple, l'expression '*E est entre A et D*' ne peut pas être formulée avec notre formalisme Ω_+ , puisque nous n'avons aucun prédicat pour dire qu'un point se trouve entre deux points.
- La présence des expressions pour lesquelles nous n'avons pas encore des règles dans notre base ERHI pour les traiter. Par exemple, l'expression '*Les extrémités du segment DE sont les milieux des côtés BC et AC*' contient deux hypothèses implicites *EstMilieuDe(D,BC)* et *EstMilieuDe(E,AC)* que nous ne pouvons pas extraire sans passer par une règle dans ERHI.
- La présence des expressions que nous n'avons pas encore vues dans l'ensemble d'entraînement. Par exemple, l'expression '*On prolonge le côté BC jusqu'en un certain point N*' contient le mot '*prolonger*', qui veut dire dans notre formalisme Ω_+

AppartientA, ne se trouve dans aucune question du corpus MSD utilisé pour l'entraînement.

6.2.2. La résolution par QED-Tutrix

Pour tester la résolution des 41 questions de test, nous utilisons le générateur de preuve QED-Tutrix. Le tableau 26 présente les résultats.

Tableau 26 : Résultats de la résolution par QED-Tutrix.

	Nombre de questions pour lesquelles nous avons trouvé au moins une solution	Nombre de questions pour lesquelles nous n'avons trouvé aucune solution
GEOSFR++ sans traitement des conjonctions de coordination	11	30
GEOSFR++ sans correction des angles	13	28
GEOSFR++	20	21

Pour GEOSFR++, le générateur de preuve de QED-Tutrix échoue à trouver les solutions pour 21 questions pour des raisons suivantes :

- Concepts et propriétés absents dans le générateur de preuve de QED-Tutrix :
 - 6 questions contiennent des concepts qui ne sont pas encore implémentés dans QED-Tutrix, qui sont : le trapèze, hexagone, obtusangle et périmètre.
 - Pour 4 questions, il manque des propriétés dans le générateur de preuve nécessaires pour générer certaines inférences présentes dans les solutions de ces questions.
- Extracteur GEOSFR++ : Pour 9 questions, l'extracteur échoue à extraire correctement les hypothèses et conclusions correctement. Voir l'analyse des erreurs dans la section 6.2.1.
- Manque d'informations dans les énoncés : Pour 2 questions, il manque des informations nécessaires à la résolution. Par exemple, pour résoudre la question : '*DEF est un triangle quelconque. La droite AB est parallèle au côté DE du triangle. Montre que*

DEF et ABF sont semblables.’, il faut préciser que *A* appartient à *FE* et *B* à *FD*. Ces informations sont présentes uniquement dans la figure de l’énoncé.

Si nous éliminons dans l’ensemble de 41 questions de test les questions pour lesquelles le générateur de preuve QED-Tutrix manque des concepts et propriétés, et aussi les questions qui ne contiennent pas toutes les informations, nous pouvons déduire que l’extracteur GEOSFR++ extrait correctement les hypothèses et les conclusions nécessaires à la résolution pour 20 questions parmi 29 questions qui restent, ce qui représente 68,96%.

Hypothèse implicite :

Dans le tableau 27, nous présentons les résultats de GEOSFR++ avec et sans le traitement des hypothèses implicites, pour ces 29 questions de l’ensemble de test. Les hypothèses implicites, ici, sont les hypothèses générées par les règles ERHI. Dans les deux cas, nous faisons le traitement des conjonctions de coordination et la correction manuelle des angles, et comparons les résultats après la conversion dans la représentation de QED-Tutrix.

Tableau 27 : Comparaison des résultats avant et après le traitement des hypothèses implicites

	Nombre de questions résolues	Pourcentage par rapport aux 29 questions
GEOSFR++ sans les hypothèses implicites	4	14%
GEOSFR++	20	69%

16 questions parmi les 20 questions que nous pouvons résoudre avec le traitement des hypothèses implicites n’ont pas pu être résolues sans faire appel à ce module, ce qui montre l’importance de ces derniers. Dans ces 16 questions, il manque majoritairement des segments, des droites, et des triangles qui ne sont pas mentionnés explicitement dans les énoncés. Mis à part cela, dans ces questions, il y a des expressions du genre ‘*La droite parallèle à la droite BC passant par A coupe la droite BI en D*’ qui nécessite une règle de ERHI pour extraire les hypothèses : *hypothese(para(line([a,d]),line([b,c])))* et *hypothese(line([b,i,d]))*. Dans cet exemple, la droite *line([a,d])* n’est pas mentionné explicitement dans l’expression.

CHAPITRE 7 DISCUSSION

Dans ce chapitre, nous présentons une discussion générale de l'ensemble des points abordés dans ce mémoire. À cette fin, nous revenons à l'objectif principal, pour voir jusqu'à quel point les solutions proposées ont pu répondre à ceux-ci. Puis, nous proposons des améliorations possibles de ces solutions.

Notre objectif principal est la création d'un extracteur automatique capable de dégager l'ensemble des hypothèses et des conclusions d'un énoncé de problème de géométrie plane, qui sont utilisées par le générateur de preuve pour la résolution du problème. Les problèmes traités sont ceux présents dans les manuels du niveau secondaire au Québec.

Dans notre recherche, nous avons essayé de créer un extracteur automatique qui soit capable d'extraire les hypothèses explicites et implicites nécessaires aux générateurs de preuve, et plus particulièrement le générateur de preuve de QED-Tutrix. À cette fin, nous avons ajouté dans notre extracteur GEOSFR++ un générateur des hypothèses implicites basé sur une liste de règles, qui va au-delà des données citées explicitement dans le texte, afin de couvrir toutes les données nécessaires à la résolution. Notons que les hypothèses explicites sont extraites à l'aide d'un classifieur entraîné par apprentissage machine. Les résultats de cette approche présentés précédemment sont satisfaisants. Cependant, GEOSFR++ a plusieurs limites :

- Notre extracteur GEOSFR++ est incapable d'extraire les connaissances des questions du type “Démontrer qu'un quadrilatère ayant 3 angles droits est un rectangle.” ou “Prouve la propriété suivante : si une droite coupe deux droites parallèles, alors les paires d'angles internes situées du même côté de la sécante sont supplémentaires.” : les classes géométriques (comme “quadrilatère”, “droite”, etc.) dans ce type de question ne sont pas nommées par des variables (comme “ABCD”, “d”, etc.). Pour traiter ce genre de question par notre extracteur, il faut assigner des variables aux classes géométriques présentes dans le texte : “Démontrer qu'un quadrilatère ABCD ayant 3 angles ABC, BCD et DAB droits est un rectangle.” et “Prouve la propriété suivante : si une droite d coupe deux droites f et g parallèles, alors les paires d'angles internes situées du même côté de la sécante d sont supplémentaires.”. Toutefois, cette assignation des variables peut être différente des choix faits par l'élève utilisant le tuteur QED-Tutrix durant sa résolution de la question, ce qui peut créer des conflits entre les deux solutions. Par

exemple, il est possible que l'élève dans l'exemple 'Démontrer qu'un quadrilatère ABCD ayant 3 angles ABC, BCD et DAB droits est un rectangle.' ait choisi des angles droits différentes que ceux fixés par l'extracteur, ce qui entraîne une mauvaise assistance de la part du tuteur à l'élève.

- Une autre limitation qui ne découle pas de notre objectif principal mais que nous jugeons importante, est l'incapacité de traiter les figures illustratives qui accompagnent les énoncés textuels dans les questions de géométrie. Nous nous sommes fixés comme objectif de faire l'extraction uniquement des questions transcrites textuellement. L'extraction des connaissances des figures s'avère très importante et incontournable, puisque certains types de données figurales ne peuvent pas être transcrites textuellement, comme l'orientation des angles, la position et l'ordre des points sur une droite qui sont nécessaires dans l'identification des angles. Ce qui nous a obligé à traiter manuellement ce genre de questions dans nos résultats. De plus, une grande partie des questions des manuels scolaires contenant des figures, qui s'avèrent être très intéressantes, sont exclues de notre recherche.
- Notre extracteur est créé pour extraire les hypothèses et les conclusions principalement pour le générateur de preuve QED-Tutrix, qui par ailleurs a été utilisé pour avoir les résultats ci-dessus. Cependant, ce générateur est toujours en cours de développement et plusieurs notions géométriques ne sont pas encore ajoutées dans ce dernier. Notons que la création de notre formalisme Ω_+ est basée sur l'analyse des manuels scolaires et non pas sur les représentations formelles des hypothèses dans QED-Tutrix. Nous avons testé notre système uniquement avec l'état actuel du générateur de preuve du QED-Tutrix, toutefois, ces tests ne sont pas parfaits, puisque nous n'avons pas les résultats de la résolution par rapport aux notions géométriques qui ne sont pas encore présentes dans QED-Tutrix.
- Une autre limitation est le nombre exponentiel des hypothèses implicites. Puisque nous ne connaissons pas encore toutes les hypothèses implicites nécessaires à QED-Tutrix, le nombre de ces dernières devient rapidement exponentiel (voir la figure 18).

le nombre d'hypothèses implicites par rapport au
nombre de points



Figure 18 : Le nombre d'hypothèses par rapport au nombre de points

CHAPITRE 8 CONCLUSION

Ce travail de recherche a apporté une contribution significative à l'extraction des connaissances des problèmes de géométrie plane, et plus particulièrement des problèmes écrits en langue française. Notre extracteur permet un encodage rapide des questions dans le tuteur QED-Tutrix, ce qui épargnera aux professeurs utilisant ce tuteur intelligent du temps à effectuer cette tâche répétitive et laborieuse.

Dans nos hypothèses de recherche, nous avons supposé qu'à partir d'un énoncé de problème de géométrie plane typique de ce qui est utilisé au Québec pour les élèves du secondaire, nous pouvons extraire automatiquement les connaissances nécessaires à la démonstration. Cependant, les résultats ont montré qu'il est insuffisant de traiter l'énoncé uniquement sans tenir compte de la figure illustrative de ce dernier. En fait, nous avons montré que pour plusieurs problèmes, il fallait faire l'extraction manuelle des angles qui ne peut se faire directement de l'énoncé. En ce qui concerne l'analyse syntaxique, nos résultats montrent l'apprentissage machine appliqué sur l'analyse syntaxique des énoncés permet d'extraire efficacement les hypothèses explicites. Toutefois, pour extraire les hypothèses implicites, nous avons dû utiliser un ensemble de règles qui permettent d'inférer ces hypothèses à partir des hypothèses implicites.

Notre travail de recherche constitue une amélioration du modèle de données mis en place par GEOS pour qu'il soit plus conforme au système d'éducation québécois, et plus particulièrement au générateur de preuve de QED-Tutrix. Malgré les limitations que nous avons signalées, l'extracteur que nous proposons est à notre connaissance le premier à traiter la langue française : il a été élaboré à la suite d'une analyse rigoureuse des questions de géométrie présentes dans neuf manuels scolaires québécois (i.e. Visions, SOMMETS, Sésamath, etc.).

BIBLIOGRAPHIE

- [1] W. Wong, S-C. Hsu, S-H. Wu, C-W Lee et W-L. Hsu. LIM-G : Learner-Initiating Instruction Model based on Cognitive Knowledge for Geometry Word Problem Comprehension. 2007.
- [2] W-L Hsu, S-H. Wu et Y-S. Chen. Event identification based on the information map-INFOMAP. 2001. Publié dans: 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236).
- [3] A. Mukherjee et U. Garain. A review of methods for automatic understanding of natural language mathematical problems. 2009.
- [4] A. Mukherjee, U. Garain et M. Nasipuri. On construction of a GeometryNet. Dans : Proceedings of IASTED international conference on artificial intelligence and applications (AIA 2007). Innsbruck, L'Autrich, 2007, p.530–536.
- [5] W. Gan et X. Yu. Automatic understanding and formalization of natural language geometry problems using syntax-semantics models. Dans : International Journal of Innovative Computing, Information and Control. Volume 14, Nombre1, February 2018.
- [6] A. Farhadi et al. M. Seo et H. Hajishirzi. Solving Gometry Problems : Combining text and diagram interpretation. Dans : Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). p.17-21, 2015.
- [7] F. Botana, M. Hohenwarter, P. Janicic et S. Weitzhofer, Automated theorem proving in GeoGebra: Current achievements, Journal of Automated Reasoning, vol.55, no.1, p.39-59, 2015.
- [8] J. Narboux. GeoProof : A user interface for formal proofs in geometry, Journal of Automated Reasoning, vol.39, no.2, p.161-180, 2007.
- [9] Z. Ye, S. C. Chou et X. S. Gao, An introduction to java geometry expert, International Workshop on Automated Deduction in Geometry, Springer Berlin Heidelberg, p.189-195, 2008.
- [10] Nicolas Leduc. QED-Tutrix: système tutoriel intelligent pour l'accompagnement des élèves en situation de résolution de problèmes de démonstration en géométrie plane. École polytechnique de Montréal, 2016.
- [11] S. C. Chou, Mechanical Geometry Theorem Proving, Springer Science & Business Media, 1988.

- [12] Y. Tian, *Everyday Practice of Plain Geometry Volume I: Foundation Part (Linear Type)*, Harbin Institute of Technology Press, 2013
- [13] H. P. Zhang, H. K. Yu, D. Y. Xiong et al.. HHMM-based Chinese lexical analyzer ICTCLAS, Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing, vol.17, p.184-187, 2003.
- [14] D. Chen and C. D Manning. A fast and accurate dependency parser using neural networks. In EMNLP. 2014.
- [15] C. Alberti, D. Andor, I. Bogatyy, M. Collins, D. Gillick, L. Kong, T. Koo, J. Ma, M. Omernick, S. Petrov et al., “Syntaxnet models for the conll 2017 shared task,” arXiv preprint arXiv:1703.04929, 2017
- [16] College Board. The college board. 2014.
- [17] M. Seo, H. Hajishirzi, A. Farhadi, et O. Etzioni. Diagram understanding in geometry questions. In AAAI. 2014.
- [18] C. Alvin, S. Gulwani, R. Majumdar, et S. Mukhopadhyay. Synthesis of geometry proof problems. In AAAI. 2014.
- [19] D. J Wales et J. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, p.101(28). 1997.
- [20] Dieter et al., A software package for sequential quadratic programming. DFVLR Obersfaffeuhofen, Allemagne. 1988.
- [21] R Cadieux et al., *Panoram@th, Cahier d'exercices. Corrigé. [1er cycle du secondaire]*, Anjou, Québec, Éditions CEC, 2011.
- [22] R Cadieux et al. *Panoram@th, Cahier d'exercices. Corrigé. [2er cycle du secondaire]*, Anjou, Québec, Éditions CEC, 2011.
- [23] C. Buzaglo et al., *Mathématiques 3000. Préparation aux examens de fin d'année. [1er cycle du secondaire]*, Montréal (Québec), Canada, 2014
- [24] Sésamath, *Sésamath Cahier CM2*, 2013.
- [25] Sésamath, *Sésamath Manuel 5e*, 2010.
- [26] Sésamath, *Sésamath Manuel 4e*, 2011.
- [27] Sésamath, *Sésamath Manuel 3e*, 2012.
- [28] L. Christine, *Perspective Mathématique. Cahier d'exercices: 1er cycle du secondaire, mathématique*, Éditions Grand Duc HRW c2005-c2006.

- [29] C. Boucher, M. Coupal, M. Jacques et L. Marotte, Intersection culture, société, et technique 3ème secondaire. Éditions Chenelière éducation.
- [30] C. Boucher, M. Coupal, M. Jacques et L. Marotte, Intersection culture, société, et technique 4ème secondaire, Éditions Chenelière éducation.
- [31] C. Boucher, M. Coupal, M. Jacques et L. Marotte, Intersection culture, société, et technique 5ème secondaire, Éditions Chenelière éducation.
- [32] C. Boucher, M. Coupal, M. Jacques et L. Marotte, Intersection sciences naturelles 4ème secondaire, Éditions Chenelière éducation.
- [33] C. Boucher, M. Coupal, M. Jacques et L. Marotte, Intersection sciences naturelles 5ème secondaire, Éditions Chenelière éducation.
- [34] C. Charron, D. Gagnon ; directeur de collection : Daniel Leclerc, Point De Vue mathématique, Séquence sciences naturelles, Cahier d'exercices. 2e cycle du secondaire. Laval, Québec: Éditions Grand Duc, 2008.
- [35] C. Charron, D. Gagnon ; directeur de collection : Daniel Leclerc, Point De Vue mathématique, Séquence technico-sciences, Cahier d'exercices. 2e cycle du secondaire. Laval, Québec: Éditions Grand Duc, 2008.
- [36] C. Boivin, D. Boivin, A. Ledoux, É. Meyer, F. Pomerleau et V. Roy, Visions. Technico-sciences. Cahier d'exercices. Corrigé. [2e cycle du secondaire]. Anjou, Québec, Les Éditions CEC, 2011.
- [37] J. R. Anderson, C. F. Boyle, et G. Yost, “The geometry tutor”, dans Proceedings of the Ninth International Joint Conference on Artificial Intelligence, vol. 1. Los Angeles, CA, USA : Morgan Kaufmann Publishers Inc., Août 1985, p. 1–7.
- [38] K. R. Koedinger et J. R. Anderson, “Abstract planning and perceptual chunks : Elements of expertise in geometry”, Cognitive Science, vol. 14, no. 4, p. 511–550, 1990
- [39] N. Balacheff, R. Caferra, M. Cerulli, N. Gaudin, M. Maracci, M. A. Mariotti, J.-P. Muller, J.-F. Nicaud, M. Occello, F. Olivero, N. Peltier, S. Pesty, S. Soury-Lavergne, R. Sutherland, J. Trgalova, et C. Webber, Baghera Assessment Project, Designing an Hybrid and Emergent Educational Society, série Les Cahiers du laboratoire Leibniz, S. Soury-Lavergne, éd. Grenoble, France : Laboratoire Leibniz-IMAG, Avr. 2003, vol. 81.
- [40] R. Gras, “Aide logicielle aux problèmes de démonstration géométriques dans l’enseignement secondaire”, Petit x, vol. 17, p. 65–83, 1988.
- [41] V. Luengo, “Cabri-Euclide : un micromonde de preuve intégrant la réfutation. principes didactiques et informatiques. réalisation”, Thèse de doctorat, Université Joseph Fourier -

Grenoble I, Grenoble, France, Sep. 1997.

[42] D. Py, “Reconnaissance de plan pour la modélisation de l’élève : le projet Mentoniez”, *Recherches en Didactique des Mathématiques*, vol. 14, no. 1.2, pp. 113–138, 1994.

[43] J. Gressier, Site web de GéométrieX, 2015. En ligne : <http://geometrix.free.fr/site/>

[44] S. El-Khoury, P. R. Richard, E. Aïmeur, et J. M. Fortuny, “Development of an intelligent tutorial system to enhance students’ mathematical competence in problem solving”, dans *Proceedings of E-Learn : World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005*, G. Richards, éd. Vancouver, Canada : Association for the Advancement of Computing in Education (AACE), Oct. 2005, pp. 2042–2049.

[45] L. Font, S. Cyr, P. R. Richard et M. Gagnon. Automating the generation of high school geometry proofs using Prolog in an educational context. 2020.

[46] Grottke, Sven et Jeschke, Sabina et Natho, Nicole et Seiler, Ruedi, mArachna: A Classification Scheme for Semantic Retrieval in eLearning Environments in *Mathematics, Recent Research Developments in Learning Technologies*, 2005.

[47] “syntaxe”. Oxford Dictionaries UK Dictionary. Oxford University Press. 2016.

[48] "Parse". dictionary.reference.com. 2010

[49] Chen, C. M. (2000). The effects of word problem represented formats and related factors on fifth graders’ problem-solving performance. Unpublished Master Thesis, National Chiayi University, Chiayi, Taiwan (in Chinese).

[50] J.C. Moyer, L. Sowder, J. Threadgill-Sowder, M.B. Moyer, Story problem formats: Draw versus verbal versus telegraphic. *Journal for Research in Mathematical Education*, 15 (5) (1984), pp. 342-351.

[51] L. Sowder, J. Threadgill-Sowder, Drawn versus verbal formats for mathematical story problems. *Journal for Research in Mathematical Education*, 13 (5) (1982), pp. 324-331

[52] Wu, J. C. (1990). The effect of diagrams for children in word problem solving. Unpublished Research Report, National Taiwan University, Taipei, Taiwan (in Chinese).

[53] S. Chou, X. Gao, J. Zhang, Automated generation of readable proofs with geometric invariants, II. Theorem proving with full-angles. *Journal Automated Reasoning* 17, 325–347 (1996)

[54] S. Chou, X. Gao, J. Zhang, A deductive database approach to automated geometry theorem proving and discovering. *Journal of Automated Reasoning* 25(3), 219–246 (2000)

- [55] S. Chou, X. Gao, J. Zhang, Automated generation of readable proofs with geometric invariants, II. Theorem proving with full-angles. *Journal Automated Reasoning* 17, 325–347 (1996)
- [56] P. Janičić, J. Narboux et P. Quaresma, The Area Method. *J Autom Reasoning* 48, 489–532 (2012).
- [57] D. Py, Environnements interactifs d'apprentissage et démonstration en géométrie, 2001.

Annexe 2 : la listes des prédicats et des fonctions de la logique Ω

Type des prédicats et des fonctions	La liste des prédicats du formalisme Ω	La liste des fonctions du formalisme Ω
unaires	Not(truth)⁶, IsCircle(circle), IsTriangle(triangle), IsQuad(quad), IsRectangle(quad), IsTrapewoid(quad), IsPolygone(polygone), IsRegular(polygone), IsAngle(angle), IsSector(sector), Equilateral(triangle), Isocetes(triangle), IsSquare(quad), IsRightTriangle(triangle), IsRightAngle(angle), IsRhombus(quad), IsArc(arc) conclusion⁷ : Find(number)	ValueOf(number)->number AreaOf(twod)->number SideOf(square)->number PerimeterOF(polygon)->number LengthOf(oned)->number SquareLength(oned)->number MeasureOf(angular)->number RadiusOf(circle)->number DiameterOF(circle)->number Sqrt(number)->number Twice(number)->number SquareOf(number)->number heightOf(polygon)->number CirconferenceOf(circle)->number HalfOf(number)->number DegreeUnit(number)->number Width(number)->number Degree()->number Pi()->number conclusion : What()->number Which()->number
binaires ou plus	Equals(number,number), Greater(number,number) IsInscribedIn(polygone,circle), IsCenterOf(point,twod) IsDiameterLineOf(line,circle), IsAreaOf(number,twod) IsLengthOf(number,oned), IsDiameterNumOf(number,circle) IsSideOf(number,quad), IsMidpointOf(point,line) IsWidthOf(number,quad), Perpendicular(line,line) IsChordOf(line,circle), Tangent(line,twod),	Point(number,number)->point Line(point,point)->line Arc(circle,point,point)->arc Circle(point,number)->line Angle(point,point,point)->angle Triangle(point,point,point)->triangle Quad(point,point,point,point)->quad Arc(circle,point,point)->arc Polygon(*point)->polygon Hexagon(6*point)->hexagon Add,Mul,Sub,Div,Pow(number,number)->number

⁶ Not ce prédicat est utilisé pour représenter la négation, par exemple 'Laquelle des réponses suivantes n'est pas vraie'.

⁷ Voir la section en bas **Représentation des conclusions.**

	Secant(line,line) IsRadiusNumOf(line,circle), IsRadiusLineOf(line,circle) PointLiesOnCircle(point,circle), PointLiesOnLine(point,line) Parallele(line,line) IntersectAt(line,line,point)	AverageOf(*number)->number RatioOf(number,number)->number SumOf(*number)->number Finalisation des relations⁸ : Is(root,root)->is CC(root,root)->cc
--	--	---

⁸ Voir la section en bas **Finalisation des relations.**

Annexe 3 : les prédicats et les fonctions de la logique Ω_+ (les concepts en gras sont les arguments présents dans les prédicats et les fonctions)

Type de la propriété	Prédicats
Avec un seul argument	root : EstNombre(root), EstBooleén(root), EstEntité(root), EstTransformation(root) entité : EstSurface(entité), EstLigneOuverte(entité), EstLigneFermée(entité), EstPoint(entité), EstEntité(entité) angle : EstAigu(angle), EstDroit(angle), EstNul(angle), EstObtus(angle), EstPlat(angle), EstPlein(angle), EstAngle(angle) ligne ouverte : EstCourbe(LO⁹), EstDroite(LO), EstSegment(LO), EstOblique(droite), EstBrisée(LO), EstLigneOuverte(LO) segment : EstDemiDroite(segment), EstArc(segment), EstRayon(segment), EstCôté(segment), EstCorde(segment), EstBissectrice(segment), EstCathète(segment), EstHypothénuse(segment), EstMédiane(segment), EstSegment(segment) droite : EstMédiatrice(droite), EstSécante(droite), EstSécante(droite), EstAxe(droite), EstTangente(droite), EstDroite(droite) ligne fermée : EstCercle(LF), EstDemiCercle(LF), EstEllipse(LF), EstPolygone(LF), EstConvexe(LF) polygone : EstDécagone(polygone), EstDodécagone(polygone), EstHeptagone(polygone), EstOctagone(polygone), EstHexagone(polygone), EstPentagone(polygone), EstHendécagone(polygone), EstQuadrilatère(polygone), EstTriangle(polygone), EstPolygone(polygone) polygone d'ordre pair : EstPolygonePair(polygone) polygone d'ordre impair : EstPolygoneImpair(polygone) triangle : EstScalène(triangle), EstIsocèle(triangle), EstÉquilatéral(triangle), EstAcutangle(triangle), EstIsoangle(triangle), EstObtusangle(triangle), EstEquiangle(triangle), EstTriangleRectangle(triangle), EstTriangle(triangle) quadrilatère : EstCerfVolant(quad¹⁰), EstCarré(quad), EstParallélogramme(quad), EstQuadRectangle(quad), EstTrapèze(quad), EstLosange(quad), EstQuadrilatère(quad) axe : EstAxeDeReflexion(axe), EstAxeDesAbscisses(axe), EstAxeDesOrdonnées(axe), EstAxeDeSymétrie(axe), EstAxe(axe) point : EstFoyer(point), EstMilieu(point), EstSommet(point), EstOrthocentre(point), EstCentre(point), EstIntersection(point), EstPoint(point) surface : EstRégion(surface), EstSecteur(surface),

⁹ Ligne ouverte comme les droites, les segments, les arcs, etc. Voir la taxonomie de la logique Ω_+ dans la Figure A.1.

¹⁰ quadrilatère;

	EstSecteurAngulaire(surface), EstSecteurCirculaire(surface), EstSurface(surface) nombre : EstMesure(nombre), EstMoyenne(nombre), EstPérimètre(nombre), EstDistance(nombre), EstLargeur(nombre), EstLongueur(nombre), EstRapport(nombre), EstAire(nombre), EstAngle(nombre), EstNombre(nombre), EstPair(nombre), EstImpair(nombre) rapport : EstRapportDeSimilitude(rapport), EstRapportDHomothétie(rapport), EstRapport(rapport) transformation : EstSimilitude(transformation), EstIsométrie(transformation), EstHomothétie(transformation), EstAgrandissement(transformation), EstRotation(transformation), EstTranslation(transformation), EstRéflexion(transformation), EstTransformation(transformation) Similitude : EstAgrandissement(similitude), EstIsométrie(similitude), EstHomothétie(similitude), EstSimilitude(similitude) Rotation : EstHoraire(rotation), EstAntiHoraire(rotation), EstRotation(rotation) Conclusions : Conclusion(hypothèse), TrouverNature(entité), TrouverValeur(nombre)
Avec deux arguments	ligne ouverte, ligne ouverte : EstMédiatriceDe(droite,segment) EstMédiatriceDeEn(droite,segment,point) EstSécanteDe(droite,LO ou LF) EstSécanteDeEn(droite,LO ou LF,point) Homologue(côté,côté) Parallèle(droite,droite) Perpendiculaire(droite,droite) CôtésAdjacents(côté,côté) Opposés(côté,côté) Congrus(côté,côté) ligne ouverte, ligne fermée : EstSécanteDe(droite,LO ou LF) EstSécanteDeEn(droite,LO ou LF,point) EstTangenteDe(droite, LF ou courbe) EstTangenteDeEn(droite,LO ou LF,point) EstMédianeDe(segment,triangle) EstRayonDe(segment,cercle) EstAxeDeSymétrieDe(axe,LF) EstAxeHorizontalDe(axe, ellipse ou losange) EstAxeVerticalDe(axe, ellipse ou losange) EstPetitAxeDe(axe, ellipse ou losange) EstCordeDe(segment,cercle) EstAxeFocalDe(axe,ellipse) EstCôtéDe(segment,polygone) EstDiagonalDe(segment,LF) EstHypothénuseDe(segment,triangle) EstCathèteDe(segment,triangle) EstHauteurDe(segment,parallélogramme ou triangle) EstRayonDe(segment,cercle) EstDiamètreDe(segment,cercle) entité , entité : Distincts(entité,entité) ligne fermée, ligne fermée : EstInscritDans(LF,LF) EstIsométrique(triangle,triangle) EstSemblable(triangle,triangle) ImageParTranslation(LF,LF) ImageParRotation(LF,LF) ImageParRéflexion(LF,LF) ImageParHomothétie(LF,LF) EstPérimètre(LF,LF) point, ligne ouverte : AppartientÀ(point, LO ou LF) EstMilieuDe(point,segment) EstExtrémitéDe(point,segment) point, ligne fermée : AppartientÀ(point, LO ou LF) EstCentreDe(point,LF) EstSommetDe(point,polygone) EstCentreDeGravité(point,LF) EstOrthocentreDe(point,triangle) EstFoyerDe(point,ellipse)

	EstIntérieurDe(point,LF) EstExtérieurDe(point,LF) EstIsocèleEn(triangle,point ou angle) EstObtusangleEn(triangle,point ou angle) EstRéctangleEn(triangle,point ou angle) nombre, ligne ouverte : EstLargeurDe(largeur,segment) EstLongueurDe(longueur,segment) nombre, ligne fermée : EstAireDe(aire,LF) EstPérimètreDe(périmètre,LF) EstCirconférenceDe(circonférence,cercle ou ellipse) aire, surface : EstAireDe(aire,surface) angle, angle : AnglesAdjacents(angle,angle) SontAlternesExternes(angle,angle) SontAlternesInternes(angle,angle) SontComplémentaire(angle,angle) SontConsécutifs(angle,angle) SontCorrespondants(angle,angle) SontOpposésParSommet(angle,angle) SontSupplémentaires(angle,angle) SontCongrus(angle,angle) SontHomologues(angle,angle) SontOpposés(angles,angles) angle, nombre : EstMesureDe(mesure,angle) angle, ligne fermée : EstAngleDe(angle,LF) EstInscritDansCercle(angle,cercle) EstCirconscriitÀ(cercle,polygone) mAnglesDuPolygone(nombre,angles,polygone) sommet,sommet : SommetsAdjacents(sommet,sommet) SommetsOpposés(sommet,sommet) nombre, nombre : DeMêmeMesure(nombre, nombre) sommet, côté : SommetOpposéÀCôté(sommet,côté) ligne ouverte, angle : EstBissectriceDe(droite,angle) BissectriceIssueDe(droite, point, polygone) MédianeIssueDe(droite, point, triangle) HauteurIssueDe(droite, point, triangle ou parallélogramme)
Avec trois arguments ou plus	rapport, nombre, nombre : EstRapportEntre(rapport,nombre, nombre) ligne fermée, ligne ouverte, point : IntersectentEn(LO ou LF, LO ou LF, point) point, point, point : SontAlignés(point, point, point) point, point, ligne ouverte : SontExtrémités(point, point, LO) nombre, point, LO, LF : EstDistanceEntre(nombre, point ou LO ou LF, point ou LO ou LF) ligne ouverte, ligne fermée, ligne fermée : EstTangenteEn(droite,cercle ou demi-cercle ou ellipse ou courbe, point) nombre, nombre, nombre : EstRapport(nombre, nombre, nombre) SommeÉgaleÀ(nombre,nombre, nombre) Conclusion(*hypothèse)
Type de la propriété	Fonctions
Avec un seul argument	ValeurDe(nombre)->nombre AireDe(LF¹¹)->aire MesureDe(Segment)->mesure MesureDe(angle)->mesure LargeurDe(LF)->largeur LongueurDe(LF ou segment)->longueur CirconférenceDe(LF)->circonférence PérimètreDe(LF)->périmètre Sqrt(nombre)->nombre Double(nombre)->nombre

¹¹ Ligne fermée comme le cercle, les polygones, les ellipses, etc. Voir la taxonomie de la logique omega+ dans la Figure A.1.

	CarréDe(nombre)->nombre moitiéDe(nombre)->nombre DegréUnité()->nombre Axe(segment)->axe
Avec deux arguments	Point(nombre,nombre)->point Ligne(point,point)->ligne Segment(point,point)->segment Arc(cercle,point,point)->arc Cercle(point,rayon)->cercle Is(root,root)->is CC(root,root)->cc Add,Mul,Sub,Div,Pow(nombre, nombre)->nombre RapportEntre(nombre,nombre)-> nombre
Avec trois arguments ou plus	Angle(point,point,point)->angle Triangle(point,point,point)->triangle MoyenneDe(*nombre)->nombre Quad(point,point,point,point)->quad Arc(cercle,point,point)->arc Polygone(*point)->polygone Hexagon(6*point)->hexagon SommeDe(*nombre)->nombre

Annexe 4 : La liste des questions pour la validation finale

Questions tirées de [57] :

Soit ABC un triangle, M le milieu du côté BC et H le milieu du segment AH. La parallèle à la droite BC passant par H coupe la droite AB en K. Montre que K est le milieu du côté AB.

Soit ABC un triangle. I est le milieu du côté AB, et J est le milieu du côté AC. La parallèle à la droite AB passant par J coupe la droite BC en K. Montre que BIJK est un parallélogramme.

Soit ABC un triangle, et H le pied de la hauteur issue de A. La parallèle à la droite BC passant par A coupe la parallèle à la droite AH passant par B en un point K. Montre que AHBK est un rectangle.

Soit ABC un triangle isocèle en A. La parallèle à la droite AB passant par C coupe la parallèle à la droite AC passant par B en un point K. Montre que ABKC est un losange.

Soit ABC un triangle. I est le milieu du segment AC, et J celui du segment AB. La droite parallèle à la droite BC passant par A coupe la droite BI en D. Montre que ABCD est un parallélogramme.

ABC est un triangle. D est le milieu du segment BC, E celui du segment AC et F celui du segment AB. I est le point d'intersection des droites AD et FE. Démontre que I est le milieu du segment EF.

ABC est un triangle isocèle en A, et M est le milieu du côté BC. Montre que MAB est un triangle rectangle en M.

Soient A, B, C, D, E, F six points placés de telle sorte que ABCD, BCED et ACFD soient des parallélogrammes. Le point O est le milieu du segment CD. Montre que ABFE est un parallélogramme.

Soit ABC un triangle. M est le milieu du côté BC. H est le point appartenant au segment AM tel que la droite BH est perpendiculaire à la droite AM. K est le point tel que M est le milieu du segment HK. Montre que KMC est un triangle rectangle en K.

Soit ABCD un quadrilatère. I est le milieu du côté AB. La parallèle à la droite BC passant par I coupe la droite AC en J. La parallèle à la droite DC passant par J coupe la droite AD en K. Montrer que la droite IK est parallèle à la droite BD.

ABC est un triangle. D, E, F sont les milieux respectifs des côtés BC, AC, AB. I et J sont les milieux des segments DE et DF. Montre que la droite IJ est parallèle à la droite BC.

Un triangle EBC est rectangle en E, et M est le milieu de l'hypoténuse. Le point A est tel que M soit le milieu du segment AE. Démontre que le quadrilatère ACEB est un rectangle.

Soit OCL un triangle. I est le milieu du côté CL. D est le point tel que ODIL est un parallélogramme. M est le milieu du segment ID. Montre que M est le milieu de du segment OC.

Soit FURS un parallélogramme. I est le milieu du côté UR. G est le point tel que I est le milieu du segment FG. Montre que R est le milieu du segment GS.

Soit ABC un triangle rectangle en A, et M le milieu du segment BC. I est le milieu du segment AB, J celui du segment AC. Montre que AIMJ est un rectangle.

Soit ABC un triangle quelconque. I est le milieu du côté AB. J est le milieu du côté AC. Les droites IC et JB se coupent en un point O. Les points M et N sont les milieux respectifs des segments OB et OC. Montre que IJNM est un parallélogramme.

Soit C un cercle de centre O et de diamètre AB. D est le cercle de diamètre AO. Q est un point de C, tel que la droite QA coupe le cercle D en un point P. Démontre que P est le milieu du segment AQ.

Questions tirées des ressources des professeurs :

Soit la droite AB perpendiculaire à la droite AE et la droite EF également perpendiculaire à AE. De plus, la droite EF est parallèle à la droite CD. Démontre que la droite AB est parallèle à la droite CD.

La droite AB coupe la droite CD en O. Si l'angle AOC vaut 142° , trouve la valeur de l'angle COB et justifie chaque étape de ta démarche.

La droite AB est coupée par la droite EC en C. La demi-droite CD forme les angles ACD et DCE qui valent respectivement 23° et 67° . Montre que les droites EC et AB sont perpendiculaires.

Soit un triangle ABC où l'angle ABC vaut 24° et l'angle BCA vaut 66° . Quelle est la nature du triangle ABC?

Le côté BC du triangle ABC correspond au diamètre du cercle C circonscrit. Démontre que le triangle ABC est nécessairement rectangle.

Soit le triangle ABC et la droite AE. L'angle EAC et l'angle BAE sont isométriques. De plus, AE coupe BC à angle droit en E. Les segments BE et CE mesurent chacun 3 centimètres. Montre que le triangle ABC est isocèle.

Dans le quadrilatère ABCD, les côtés AD et BC sont parallèles et l'angle BAD vaut 146° . On prolonge le côté BC partir du sommet C. On place le point E de sorte que l'angle CDE vaut 88° et l'angle CED 58° . Montre que le quadrilatère ABCD est un parallélogramme.

Soit le cercle C_1 de centre A et de rayons AB et AD. On construit le cercle C_2 de centre B et de rayon AB et le cercle C_3 de centre D et de rayon AD. De plus, l'angle DAB est droit. Le point C est placé à l'intersection des cercles C_2 et C_3 . Le segment b_1 est supporté par la bissectrice de l'angle ABC. Trouve la valeur de l'angle CBE.

Soit le triangle équilatéral ABC et E, F et G respectivement les points milieux des côtés AB, BC et AC. Démontre que EFG est équilatéral.

Dans un losange ABCD, démontrent que la diagonale AC est la bissectrice de l'angle BAD.

Soit le carré ABCD et E, F, G et H respectivement les points milieux des côtés AB, BC, CD et AD. Montre que EFGH est aussi un carré.

DEF est un triangle quelconque. La droite AB est parallèle au côté DE du triangle. Montre que DEF et ABF sont semblables.

Les extrémités du segment DE sont les milieux des côtés BC et AC du triangle ABC. Montre que DE est parallèle à AB.

Les extrémités du segment DE, qui mesure 10 cm, sont les milieux des côtés BC et AC du triangle ABC. Montre que la mesure de DE est la moitié de celle de AB.

Soit un triangle quelconque ABC dont l'angle B est aigu. On prolonge le côté BC jusqu'en un certain point N et on mène la bissectrice CM de l'angle ACM . Puis, par le point B , on mène une parallèle à cette bissectrice qui rencontre CA en E . Prouver que $m\ BC = m\ CE$.