**UNIVERSITÉ PARIS 8 – VINCENNES – SAINT-DENIS**
**ECOLE DOCTORALE COGNITION, LANGAGE ET INTERACTION**

**THESE**

Pour obtenir le grade de

**DOCTEUR DE L'UNIVERSITE PARIS 8**

Discipline : Psychologie Cognitive

Par

Bruno MARTIN

Le 11 juillet 2016

---

# *DIAGNOSTIC COMPORTEMENTAL ET COGNITIF DES ERREURS DANS LA RESOLUTION DE PROBLEMES ARITHMETIQUES*

---

Devant le jury composé de :

| | | |
|---|---|---|
| Jacques JUHEL | Professeur, Université de Rennes 2 | (Rapporteur) |
| André TRICOT | Professeur, Université de Toulouse | (Rapporteur) |
| Jean-François RICHARD | Professeur Honoraire, Université Paris 8 | (Invité) |
| Didier BAZALGETTE | Direction Générale de l'Armement | (Invité) |
| Jean-Marc LABAT | Professeur, Université Paris 6 | (Co-encadrant) |
| Emmanuel SANDER | Professeur, Université Paris 8 | (Directeur) |

# REMERCIEMENTS

En premier lieu, je tiens à remercier mon directeur de thèse Mr Emmanuel Sander pour sa confiance, sa patience, sa rigueur intellectuelle hautement formatrice et surtout pour m'avoir proposé de travailler sur ce sujet de recherche passionnant.

Merci à mon codirecteur de thèse Mr Jean-Marc Labat, pour sa bonne humeur communicative, sa bienveillance et la pertinence de ses conseils. Il m'aurait été difficile de mener correctement l'aspect EIAH de ma thèse sans son soutien actif.

Un immense merci à Mr Jean-François Richard à qui je dois beaucoup d'un point de vue intellectuel. Nos échanges étaient des plus enrichissants et je lui dois les idées principales développées dans ma thèse. Je le remercie pour sa passion contagieuse qui m'a conduit à la conviction de l'utilité théorique et applicative de la modélisation cognitive en psychologie.

Merci à Mr André Tricot et à Mr Jacques Juhel d'avoir accepté d'être les rapporteurs de ce travail, et merci aux membres du jury pour leur présence.

Merci à Valentine, je lui suis extrêmement redevable. Je n'ose imaginer ce que ma thèse aurait pu être sans ses données. Merci aussi pour son soutien et ses importants travaux de relecture.

Un grand merci à Ysoline, ma conjointe, pour son aide et son support moral inconditionnel. Merci pour sa patience indéfectible, pourtant mise à rude épreuve au cours de cette thèse, et plus particulièrement dans la relecture de la première version de ce document.

Merci à mon collègue de galère, Timothée, pour son amitié, son support, et son aide cruciale dans le développement de DIANE.

J'exprime aussi ma reconnaissance à Khider, mon prédécesseur, pour avoir lancé le projet DIANE. Sans son travail je n'aurais pas pu effectué le mien.

Merci plus généralement à tous les membres de l'équipe CRAC qui m'ont permis de baigner dans un environnement humainement et intellectuellement riche.

Je ne sais comment exprimer ma gratitude aux membres de l'équipe MOCAH qui m'ont accueilli comme l'un des leurs d'abord en tant qu'ATER puis en tant qu'ingénieur. Merci donc pour leur confiance et pour l'ambiance stimulante qui règne dans leur équipe. Merci tout particulièrement à Mathieu et à Amel, pour m'avoir permis de

travailler parmi eux tout en me laissant suffisamment de flexibilité pour finaliser ma thèse.

Merci à mes parents et plus largement à mes proches et mes amis pour leur soutien et leur compréhension.

Je n'oublie pas d'adresser mes remerciements aux élèves de l'école primaire de Veynes qui ont vaillamment résisté à la torture de mes problèmes de billes, ainsi qu'aux professeurs m'ayant permis de réaliser cette expérimentation.

Je remercie finalement la Direction Générale de l'Armement, qui a financé cette thèse, et en particulier mon correspondant, Mr Didier Bazalgette.

# Table des matières

# GLOSSAIRE ET ACRONYMES PRINCIPAUX

- ACT-R : Active Control of Thought. Architecture cognitive centrée sur les règles de production (si<condition>alors<action>) développée par Anderson.
- CBT : Constraint Based Tutor. Tuteur basé sur les contraintes. Type de tuteur centré sur la détection de viol de contraintes qui décrivent les connaissances du domaine.
- Diagnostic comportemental : découverte du cheminement de l'élève conduisant à la réponse qu'il propose.
- Diagnostic épistémique : inférence sur les connaissances ou tout autre trait cognitif sur l'élève
- Diagnostic cognitif : composition du diagnostic comportemental et du diagnostic épistémique. Étant donné que le processus de diagnostic se « conclut » sur le diagnostic épistémique, nous effectuons parfois l'abus de langage consistant à désigner le diagnostic épistémique par « diagnostic cognitif ».
- EIAH : Environnement Informatique pour l'Apprentissage Humain
- DIANE : EIAH dont l'acronyme signifie Diagnostic Informatique pour l'Arithmétique en classe Elémentaire.
- IRT : Item Response Theory. Technique issue de la psychométrie basée sur des régressions logistiques pour établir un diagnostic du sujet.
- MDL : Minimum Description Length. Principe issu de la théorie de l'information indiquant que si un modèle découvre de la régularité dans des données, alors il peut compresser ces dernières.
- PAEV : problème arithmétique à énoncé verbal.
- Problème complexe : problème à plusieurs étapes. Dans cette thèse, ce terme correspond seulement à un cas particulier de problèmes à plusieurs étapes dont la résolution peut avoir lieu en une ou trois opérations.
- RSM : Rule Space Method. Méthode de diagnostic développée par Tatsuoka construite au-dessus de l'Item Response Theory.
- STAR : Simple Toolbox to Analyse Reasonning. Programme développé durant la thèse pour exprimer et tester des modèles cognitifs symboliques simples.

# 1 INTRODUCTION

## 1.1 Description du problème

> *Simon avait des billes avant la récréation. Il gagne 6 billes et maintenant a 13 billes. Combien Simon avait-il de billes ?*

Ci-dessus est représenté un exemple classique de problème à énoncé verbal. Les premières années en classe élémentaire forment le tout premier contact de l'élève avec la résolution de problème mathématique décrit par un texte. Ils décrivent une situation dans laquelle une quantité doit être déduite à partir des informations données dans le texte et de leurs connaissances mathématiques. Notre thèse portant sur l'étude de leur résolution, nous proposons l'acronyme **PAEV** pour signifier **Problème à Énoncé Verbal**.

> *« Through problem solving, students can experience the power and utility of mathematics. Problem solving is central to inquiry and application and should be interwoven throughout the mathematics curriculum to provide a context for learning and applying mathematical ideas »*

Cette citation, extraite des recommandations National Concil of Teachers Of Mathematics (NCTM, 2000, p. 256), met l'accent sur une dimension importante de la résolution de problème[1]. Elle la place comme une activité tournée vers l'extérieur,

[1] Les États-Unis n'ont pas de programme national de mathématiques. Pour lutter contre le faible niveau en mathématiques comparé à d'autres pays développés, le National Concil of Teachers Of Mathematics a

justifiant l'intérêt des apprentissages mathématiques en leur donnant un sens. Plus particulièrement dans le cadre des PAEV, ce sont les opérations mathématiques telles l'addition et la soustraction qui prennent du sens. En effet, il est maintenant connu que l'aspect procédural des opérations mathématiques n'est pas dissocié de ses aspects sémantiques, c'est-à-dire les situations dans lesquelles elles peuvent être effectuées (Brissiaud & Sander, 2010). Enseigner aux élèves la résolution de problème arithmétique à énoncé verbal est à la fois une fin, en permettant la résolution de problèmes réalistes et un moyen dans la mesure où connecter l'opération à son sens est nécessaire pour sa maitrise. Les connaissances mathématiques sont généralement organisées de manière hiérarchique, il est donc crucial de maîtriser certaines compétences avant d'en aborder d'autres. Ainsi, la résolution de problèmes d'algèbre dans lesquels le sujet doit mettre en équation des relations mathématiques avec une ou plusieurs inconnues ne peut se passer d'une bonne maitrise des PAEV. Ces premiers contacts avec les mathématiques sont donc cruciaux. La résolution de PAEV peut être analysée sous divers angles dont voici une liste non exhaustive tant dans les approches que dans les références.

- Les procédures de calcul et les aspects numériques (Daroczy, Wolska, Meurers, & Nuerk, 2015; Dehaene, 2001).
- La charge en mémoire de travail (LeBlanc & Weber-Russell, 1996).
- Les schémas et les connaissances conceptuelles (Kintsch & Greeno, 1985; Marshall, 1995).
- La représentation de la situation et les approches incarnées (Brissiaud & Sander, 2010; Reusser, 1990).
- L'encodage des relations (Chaillet, 2014; Sander & Richard, 2005).
- La lecture et la compréhension (Cummins, Kintsch, Reusser, & Weimer, 1988).

---

été créé pour émettre des recommandations. En 1989, il souligne l'importance de l'activité de résolution de problème et insiste en 2000 sur ce point.

- L'impact de l'artificialité des problèmes sur le raisonnement (De Corte, Verschaffel, & Greer, 2000; Gerofsky, 2010; Reusser & Stebler, 1997).
- L'impact des mots-clefs sur la résolution (Hegarty, Mayer, & Monk, 1995; Perla Nesher & Teubal, 1975).

Cette diversité d'approches rend l'étude des PAEV riche, mais problématique. La pluralité des compétences, stratégies et difficultés des élèves fait de la **modélisation** un processus difficile. Elle est aussi cruciale dans la mesure où il est souhaitable d'unifier ou de comparer des approches plutôt que les étudier de manière compartimentée.

> *Modeling is a principal – perhaps the primary – tool for studying the behavior of large complex systems*

Cette citation de Simon (1990, p. 7) qui est un père de la modélisation cognitive, défend ainsi la modélisation comme outil fondamental pour l'étude de la cognition humaine. Newell, son principal collaborateur, tient une position similaire dans un article (Newell, 1973) intitulé *« You can't play 20 questions with nature and win[...] »*. Il soutient que les recherches expérimentales en sciences cognitives, compte tenu de leur impressionnante floraison, doivent absolument être contrebalancé par des démarches théoriques et modélisatrices dans le but de « mettre ces phénomènes ensemble » *(« putting them together »)*

> *There is, today, an amazing number of such phenomena that we deal with. The number is so large it scares me.(p. 2)*

Plus loin, dans la conclusion :

> *Maybe we are reaching the day of the theorist in psychology, much as it exists in other sciences such as physics. Then the task of putting things together falls to them, and experimentalists can proceed their own way.(p. 24)*

Le jour des « théoriciens » s'est-il produit ? Il nous est bien sûr difficile de tenir des affirmations générales sur la forme actuelle des recherches en psychologie cognitive.

Dans le domaine des PAEV, toutefois, il est apparent que les approches modélisatrices, qui ont connu un âge d'or dans les années 80, se font relativement rares aujourd'hui[2].

Mettons de côté un instant la psychologie cognitive. Dans la recherche sur les **E**nvironnements **I**nformatique pour l'**A**pprentissage **H**umain (EIAH), la modélisation du comportement humain est une problématique très vive. Le terme employé n'est pas « modélisation du sujet », mais « modélisation de l'apprenant ». Le terme anglophone, « Student Modeling », donne en mars 2016 plus de 8000 résultats sur Google Scholar[3]. La modélisation de l'apprenant est particulièrement importante dans le domaine des « Tuteurs Intelligents »[4]. Leur but est d'optimiser les apprentissages, en visant un rapprochement si ce n'est un dépassement de la qualité de l'instruction avec un tuteur humain expert dans des interactions du type seul à seul. Ce type de tutorat est connu pour donner de très bonnes progressions scolaires. Sur la base de cette observation, Bloom (1984, p. 15) pose le problème suivant : « *find methods of group instruction as effective as one-to-one tutoring* », défi que reprend à son compte la recherche sur les tuteurs intelligents (Corbett, 2001). Un de leur enjeu fondamental est donc la création et l'utilisation de modèles pertinents de l'apprenant (« *student modeling* »). Cette préoccupation les distingue donc des systèmes d'enseignement assisté par ordinateur développés au début des années 1970 qui n'offraient qu'une personnalisation faible des apprentissages.

Devant la convergence d'intérêts entre la psychologie cognitive et les environnements d'apprentissage, certains chercheurs ont considéré qu'un mariage était sur le point de se produire. Dans les années 80, la relation entre les tuteurs intelligents et la psychologie cognitive était vue par Anderson comme symbiotique (J. Anderson, 1984 cité dans

[2] À l'exception, au moins, de (LeBlanc & Weber-Russell, 1996), mais dont le modèle est centré sur la modélisation de la demande en mémoire de travail des PAEV.

[3] Sauvegarde de la recherche : http://archive.is/mo145

[4] Nous soulignons que notre thèse porte plutôt sur les systèmes qui mettent de côté la problématique du tutorat au profit de la question du diagnostic. Toutefois, l'essentiel des travaux en modélisation de l'apprenant ont été réalisés avec la motivation d'offrir un tutorat de l'apprenant, c'est pourquoi nous analysons la littérature sous l'angle des tuteurs intelligents.

Wenger, 1987). Ohlsson, chercheur issu de la psychologie cognitive et très prolifique dans le domaine des environnements d'apprentissage, a aussi tenu une position forte[5] (Ohlsson, 1991) suggérant que le champ « *éducation et intelligence artificielle* » devrait, à terme, devenir une branche de la psychologie cognitive du fait de sa capacité et son intérêt à tester des théories de l'apprentissage. Cet idéal est attrayant. Dans notre domaine d'étude qui sont les PAEV, une dynamique de recherche couplée peut alors être imaginée : la mise en place de modèle cognitif est facilitée par l'environnement informatique par les données qu'il procure, et en retour le modèle cognitif peut améliorer les apprentissages au sein de l'environnement en les personnalisant.

Or, l'analyse des techniques actuelles de modélisation de l'apprenant remet en doute que l'idéal suggéré par ces chercheurs ait eu lieu. En effet, la profondeur psychologique des modèles mis en place dans les environnements d'apprentissage est généralement faible. Nous justifierons ce point au travers du manuscrit. Dans l'approche dominante actuelle, un modèle de l'apprenant est généralement une forme normative des connaissances que le sujet doit obtenir, parfois complémentée par quelques connaissances erronées.

Le diagnostic cognitif est donc une notion qui a de multiples facettes. Ohlsson et Langley (1988, p. 42) l'illustrent explicitement ici :

> *To the psychologist the problem of cognitive diagnosis is the problem of "research method" - how to process empirical observations of subjects. To the teacher it is the problem of "assessment" - how to evaluate the knowledge, or lack of knowledge, of students. To the computer scientist it is part of the problem of "user modeling" - how to construct interfaces that adapt to the individual user.*

L'idée majeure des tuteurs d'interagir au plus vite avec l'apprenant peut être défendue pour des raisons d'efficacité (Corbett & Anderson, 2001), mais limite forcément la

[5] La position de Ohlsson était peut-être provocatrice, car écrite dans le cadre d'une section libre nommée « viewpoint »

profondeur cognitive du diagnostic. Il est difficile de prévoir l'usage que peut avoir un modèle cognitif tant qu'il n'a pas fait ses preuves. En ce sens, nous notons que la règle de Self (1990, p. 11) (« *don't diagnose what you can't treat* ») peut avoir un effet « *frein* ». En effet, comment innover dans le tutorat de l'apprenant si les démarches de recherches portant sur la modélisation du sujet sont trop conservatrices et qu'elles n'apportent pas de nouvelles connaissances sur ce qui compose ses difficultés ou ses stratégies ? Si l'effort de modélisation cognitive n'est pas réalisé, alors les nouvelles formes de connaissance sur le sujet ne peuvent émerger, et les nouvelles formes d'interactions basées sur ces connaissances ne peuvent être imaginées. Nous soumettons l'idée que le relâchement de la contrainte consistant à construire des interactions pédagogiques avec l'enfant permet au chercheur d'aborder une démarche plus fine et plus fondamentale d'un point de vue théorique pour la modélisation et le diagnostic. Nous nommons dans la suite « **systèmes de diagnostic** » les outils qui se centrent sur la production de diagnostic sans chercher à appuyer l'apprenant dans sa résolution de problèmes. En référence à la citation de Ohlsson ci-dessus, c'est le point de vue cognitif que nous prenons en priorité tout en essayant de développer une approche qui puisse avoir des perspectives intéressantes en EIAH et ainsi de se rapprocher des deux autres dimensions pointées par Ohlsson : l'adaptation ou l'évaluation (assessment). Dans ces outils, un usage pédagogique du diagnostic reste possible. Il peut être fourni au professeur pour connaître les difficultés de sa classe ou d'élèves en particulier, aider à la constitution de groupes dans le cadre d'activités de remédiations, ou être utilisé à même l'outil pour lancer des programmes d'apprentissage sur mesure.

Notre problématique directrice est la suivante : comment construire une modélisation de l'apprenant capable de tirer des diagnostics cognitifs attentive aux productions d'un sujet dans leur totalité (plusieurs réponses analysées par sujet) et dans leur spécificité (bon niveau de détail dans les données) ? Pour répondre à cette problématique, nous nous limitons au cadre de la résolution de problèmes arithmétiques en classe élémentaire tout en essayant de développer des approches génériques.

## 1.2 Vue d'ensemble du document

Outre la problématique et la conclusion, notre manuscrit se décompose en 7 chapitres. Quatre sont constitutifs de l'état de l'art et trois décrivent nos contributions.

Le premier chapitre présente un état de l'art des techniques de modélisation de l'apprenant présents dans la recherche en EIAH. Une analyse des aspects cognitifs de

ces approches est présentée. Le but sera d'évaluer la proximité de la psychologie cognitive avec les environnements d'apprentissage et d'examiner la nature de leur relation. La réflexion est d'abord générale, puis se porte sur 6 exemples de systèmes. La présentation de ces outils ne permet pas seulement d'exemplifier nos analyses, mais est utilisée plus tard car ils sont rediscutés plus loin dans la thèse dans des thématiques différentes.

Le deuxième chapitre porte sur les techniques d'évaluation des diagnostics cognitifs. Nous montrerons que les techniques d'analyse des modèles qui nous semblent les plus pertinents pour des approches cognitives fines posent problème. Les exemples analysés dans le chapitre précédent sont à nouveau étudiés à la lumière de cette problématique d'évaluation statistique. Enfin nous décrivons le principe de minimum de complexité et son emploi actuel dans l'évaluation des modèles.

Le troisième chapitre porte sur les sources d'erreurs dans la résolution de problème arithmétique. Ce chapitre est mono disciplinaire et propose une analyse de la littérature sur le domaine des PAEV, guidée par le souhait de déterminer quels sont les processus cognitifs sous-jacents à leurs résolutions justes ou erronées.

Le quatrième et dernier chapitre de notre état de l'art propose une revue des différents EIAH dans le domaine des PAEV. Cet état de l'art permet de préciser nos analyses sur la profondeur de la modélisation de l'apprenant dans notre domaine d'étude. Nous montrons que les modélisations comportementales et épistémiques de l'apprenant sont minimales et que les connaissances dont le troisième chapitre fait état ne sont pas exploitées dans ces environnements.

Nos chapitres de contributions sont au nombre de trois. Dans le premier nous décrivons le module de diagnostic comportemental mis en place dans l'EIAH DIANE. Le module de diagnostic comportemental a été construit pour être évaluable, ce qui fait l'originalité de l'approche dans un cadre d'une recherche croisée avec la psychologie cognitive. Nous décrivons sa méthode de conception et les analyses employées pour étudier sa pertinence.

Dans le deuxième chapitre, nous développons et évaluons des modèles cherchant à expliquer les erreurs dans le PAEV. Ce chapitre utilise et implémente les processus que le quatrième chapitre de l'état de l'art a identifiés. Des modèles exécutables sont utilisés pour générer des erreurs dans la résolution de problèmes avec de multiples étapes. Ces

erreurs générées sont comparées aux erreurs réelles des élèves et permettent ainsi d'évaluer et de comparer les modèles mis en place.

Dans le troisième chapitre, nous décrivons et utilisons une méthode construite pour permettre d'évaluer la pertinence de modèles symboliques déterministes basée sur le minimum de complexité. À l'aide d'un logiciel développé pour écrire et évaluer des systèmes de diagnostic par cette méthode, un modèle de diagnostic est testé. Ce chapitre emploie le diagnostic comportemental développé dans la première contribution et les connaissances développées dans la deuxième.

# PREMIERE PARTIE

# ÉTAT DE L'ART

# 2 PROFONDEUR COGNITIVE DES EIAH

Nous dressons dans ce chapitre un court état de l'art des techniques de modélisation employées dans les environnements en étudiant par la suite au lien transdisciplinaire entretenu avec la psychologie dans le domaine en général et avec ces différentes techniques. Nous abordons ensuite l'angle de la modélisation des erreurs en psychologie.

Nous commençons par dresser une vue d'ensemble des différents techniques portant sur la modélisation de l'apprenant, suivi d'une description plus détaillée de quatre techniques choisies pour leurs aspects cognitifs. Ces approches seront reprises dans la suite du document par un questionnement critique sur la valeur qu'ils peuvent avoir dans un contexte de recherche en psychologie cognitive. Ces exemples seront repris lorsque nous nous poserons la question dans le chapitre suivante de la validation statistique des diagnostics établis en s'intéressant particulièrement au problème des paramètres libres et du risque de surajustement.

## 2.1 Modélisation de l'apprenant : un court état de l'art des techniques employées

Selon les auteurs, les définitions varient pour un même terme et peuvent au contraire être similaires pour des termes différents. Par exemple pour de nombreux auteurs, le terme « *student modeling* » est très proche des termes *« cognitive diagnostic »* (diagnostic cognitif) ou *« student diagnosis »* (Dillenbourg & Self, 1992; Ragnemalm,

1995; VanLehn, 1988; Wenger, 1987). En témoigne, par exemple, la proximité de ces deux définitions proposées dans la littérature.

Ragnemalm (1995, p. 1) :

> *In this paper **student diagnosis** is defined as the abstract process of gathering information about the student and turning that information into the basis for instructional decisions made in the ITS.*

Chrysafiadi & Virvou (2013, p. 4716):

> ***Student modeling** can be defined as the process of gathering relevant information in order to infer the current cognitive state of the student, and to represent it so as to be accessible and useful to the tutoring system for offering adaptation*

La mise en gras est la nôtre. Lorsqu'une différentiation est faite, il est souvent proposé que le diagnostic cognitif est le processus qui, en cours de session d'apprentissage, constitue et met à jour le modèle de l'élève (Ragnemalm, 1995; VanLehn, 1988). Le terme « *student modeling* », plutôt consensuel, ne correspond pas à la modélisation du comportement général de l'étudiant, mais à la découverte du modèle **d'un apprenant à un moment donné**. La définition de « *cognitive diagnosis* » est plus ou moins restreinte selon le programme de recherche porté par ces derniers. Pour les psychométriciens, le diagnostic cognitif correspond à la recherche de traits latents représentant des compétences (de la Torre, 2009). Selon Self (1993), Wenger (1987) et VanLenh (1988), il correspond plutôt à la recherche d'états cognitifs du sujet à partir de ses performances. La définition la plus large est alors celle de Wenger (p. 367) qui parle d'information en général :

> *The word diagnosis has been used to refer to pedagogical activities aiming at collecting and inferring information about the student or his actions.*

Dans ce cas, la différentiation avec le terme « student model » prend une autre dimension. En effet, Wenger (1987, p. 367) écrit ensuite :

> *Because this task often involves the construction of a student model, these activities have also been called student modeling. In this chapter, we adhere to the more general term diagnosis, because some of the activities we describe do not result in what is usually called a student model.*

La même idée peut être retrouvée dans Ohlsson (1986, p. 8) dans une note en bas de page :

> *In research on tutoring system, the term "student modelling" is often used, but I will avoid it, since it is unnecessarily restrictive. Students are not the only human beings who can be diagnosed, and a model is not the only possible kind of diagnostic.*

Cette distinction est intéressante pour éviter de perdre en généralité, mais notre approche étant centrée sur le diagnostic construit à partir de modèles cognitifs, nous laissons de côté cette proposition. Un bref état de l'art de ces techniques est présenté selon trois axes principaux: (1) La forme et le contenu des connaissances attribuables à l'apprenant, (2) la manière dont elles sont attribuées à l'élève en cours de session (le diagnostic) et (3) l'usage pédagogique ultérieur de cette attribution.

Les connaissances attribuables à l'élève sont généralement décrites en prenant pour point de référence l'expertise dans la résolution. À partir de ce modèle expert, celui de l'apprenant peut être construit. La première possibilité est de considérer que les connaissances de l'élève constituent un sous-ensemble des connaissances de l'expert. C'est l'approche la plus répandue, elle porte le nom d'« *Overlay* » (modèle de recouvrement partiel).

La deuxième possibilité est d'inclure des connaissances fautives (retrouvées parfois sur le terme de « *bug* » ou de « *misconception* » selon le caractère respectivement procédural ou conceptuel de la connaissance fautive mise en jeu). Elles sont généralement représentées sous la forme d'un catalogue, adjoint aux connaissances expertes pour pouvoir compléter l'ensemble des connaissances possibles de l'élève. Il s'agit alors d'un « perturbation model ». Leur niveau de granularité est variable. Il existe des modèles de types « *glass box* » (boîte transparente) et de type « *black box* » (boîte noire). Ces derniers sont les plus simples, car ils portent simplement sur l'état final lors de la résolution de problème. Les premiers sont plus ambitieux et cherchent à modéliser la résolution pas-à-pas en considérant les états intermédiaires. Ils nécessitent des outils de diagnostics comportementaux permettant de retracer les opérations de l'apprenant, comme décrit dans le paragraphe suivant.

La forme des connaissances peut ainsi varier selon les outils (Woolf, 2009, p. 56) : « *procédurale* » (règles d'actions, mais aussi plans, buts et tâches) ou « déclarative » (généralement héritée de la logique propositionnelle). L'écriture des connaissances dans

les systèmes intelligents comporte aussi de nombreuses alternatives allant de l'écriture de règle de productions (« *si <condition> alors <action>* ») aux ontologies en passant par la logique formelle et les réseaux sémantiques.

Les techniques de diagnostic peuvent être classées en deux grandes catégories : les « *symboliques* » et « *numériques* ». Les premières considèrent l'aspect « algorithmique » de l'élève. Elles sont généralement employées pour analyser en profondeur des réponses, en étudiant les pas successifs réalisés dans la résolution. De multiples approches ont été proposées pour franchir le pont entre les données directement accessibles dans l'environnement et le diagnostic. Lorsque le modèle de l'apprenant décrit une suite d'opération permettant de résoudre un problème, il est utile d'inférer à partir des données les états intermédiaires du problème voire des états mentaux traduisant les représentations successives que le sujet se fait de l'état du problème (VanLehn, 1988). Il faut donc un pont[6] entre les données en entrée et l'attribution de connaissances (ou autres informations) à l'élève. Diverses techniques ont été proposées pour tenter de remplir cet objectif. Le Model Tracing (traduit parfois par traçage de modèle) en est un exemple. Il dérive de la théorie cognitive ACT (et successeurs) qui propose que les compétences cognitives soient représentées comme des règles de production. Il simule une résolution de problème et permet donc de faire un lien entre les règles employées par l'élève et ses états mentaux. Ces derniers ne sont pas des données directement accessibles, d'autres techniques permettent donc de complémenter le Model Tracing comme la découverte de chemin (inférer des états mentaux à partir de l'état final) ou la reconnaissance de plan (inférer des états mentaux à partir d'états intermédiaires). Dans le cadre des modèles qui nécessitent ce type d'inférence, la modélisation de l'apprenant est divisée en deux parties : une modélisation **comportementale** décrivant ce qu'il a fait, et **épistémique** donnant une signification à ces comportements, notamment en lui attribuant des connaissances (ou

[6] Le terme « pont » (bridge) est emprunté à Ragnemalm (1995) qui propose une revue plus complète que notre résumé en catégorisant et représentant graphiquement les différentes approches diagnostiques selon les étapes et les niveaux de granularités caractérisant ces approches.

des conceptions erronées) (N. Balacheff, 1994; Wenger, 1987)[7]. Selon ces deux chercheurs, le diagnostic épistémique (les connaissances qui lui sont attribuées) de l'apprenant dépend fortement de la fidélité et de la richesse de sa modélisation comportementale. VanLehn utilise le terme largeur de bande (« *bandwidth* ») pour qualifier la richesse des données fournies au programme réalisant le diagnostic épistémique[8]. Les méthodes numériques, à la différence des méthodes symboliques, sont utilisées pour raisonner dans l'incertitude dans l'attribution d'une connaissance à un apprenant. Le principe est de quitter la binarité (connaissance obtenue ou non), du fait du caractère non déterministe des comportements. En effet, malgré la maitrise d'un objet de connaissance, des erreurs d'exécutions peuvent se produire, par inattention par exemple. Des techniques d'apprentissage statistiques (Machine Learning) sont alors utilisées pour associer une probabilité à l'obtention des connaissances ou à la réussite ou non d'un problème. Elles sont actualisées au cours de la session d'apprentissage. Il en existe plusieurs, notamment les régressions logistiques. Elles cherchent à estimer la probabilité que l'apprenant *i* aie la connaissance *j*. Deux grandes familles sont utilisées : Les Learning Factor Analysis (LFA) et l'Item Response Theory (IRT). Les LFA sont issues de l'Educational Data Mining et IRT de la psychométrie. À l'intérieur de chaque famille, des variations existent selon les paramètres introduits pour estimer la probabilité qu'un élève possède une connaissance. D'autres méthodes numériques existent comme les chaînes de Markov ou les réseaux bayésiens. Lorsque cette attribution a lieu dans le cadre d'un tuteur intelligent, ces méthodes sont parfois employées sous le nom de Knowledge Tracing[9] (traçage de connaissance).

---

[7] Balacheff (1994) reprend à son compte la distinction de Wenger, mais propose une variation en suggérant que les inobservables comme les intentions ou les buts doivent se situer au niveau épistémique dans la mesure où ils sont inférés par un système.

[8] VanLehn (1988) n'emploie pas le terme de diagnostic épistémique, seulement le terme « diagnostic ». Cependant son usage du terme correspond à la notion de diagnostic épistémique chez Wenger, d'où notre raccourci de langage.

[9] À l'origine, le Knowledge Tracing (KT) est porté par une approche bayésienne et permettait de compléter le Model Tracing (Corbett & Anderson, 1994). Or, depuis, des alternatives ont été proposées.

Le diagnostic peut intervenir de plusieurs manières dans les environnements. Dans les tuteurs intelligents, un système de feedback est généralement mis en place dans les environnements. Il peut être de formes diverses (Corbett & Anderson, 2001) : immédiat, indication de présence d'erreur seulement, aide sur demande. Certains systèmes offrent la possibilité à l'apprenant d'observer, voire de corriger le modèle représentant leurs compétences (Joséphine, Nkambou, & Bourdeau, 2006), il s'agit alors d'un « modèle ouvert ». Certains outils ne cherchent pas à tutorer et ne fournissent pas de feedbacks ni d'indices lors de la résolution. Ces environnements peuvent être décomposés en deux catégories selon s'ils disposent ou non d'un module de diagnostic. Dans les outils sans diagnostic, la richesse de l'activité supportée par le système l'emporte sur la nécessité de feedbacks. En font partie les outils d'inspirations constructivistes comme les micro-mondes (N. Balacheff, 1994) qui laissent l'enfant effectuer des manipulations d'objets virtuels avec l'espoir qu'il construise son savoir par lui-même. Les outils avec module de diagnostic permettent généralement de produire un compte rendu à l'usage du professeur décrivant le profil de ses élèves (El-Kechaï, Delozanne, Prévit, Grugeon, & Chenevotot, 2011; Hakem, Sander, Labat, & Richard, 2005). Nous nommons cette catégorie d'outils « *systèmes de diagnostic* ».

Les trois dimensions constituant cet état de l'art ne sont pas indépendantes. Lorsque les connaissances sont décrites à un niveau grossier, les techniques symboliques permettant de retracer le processus de solution de l'élève ou la détection de « bug » ne sont pas nécessaires. Selon que l'outil est un tuteur intelligent ou un système de diagnostic, les approches peuvent différer. Par exemple, dans la famille des approches numériques, l'IRT est généralement utilisée pour analyser le résultat hors-ligne de batteries de tests tandis que les approches de type Knowledge Tracing sont plus adaptées (et donc plus utilisées) pour actualiser les probabilités au cours de session d'apprentissage avec un Tuteur Intelligent. En effet, l'IRT, dans sa formulation classique, ne prend pas en compte la progression de l'apprenant, la rendant peu adaptée aux systèmes visant la

---

Dans certaines publications, ces approches gardent le terme (KT) et l'approche bayésienne est nommée BKT (e.g d Baker, Pardos, Gowda, Nooraei, & Heffernan, 2011)

progression de l'apprenant. Les deux approches ont donc une différence de point de vue sur l'usage pédagogique du diagnostic.

## 2.2 Liens entre les disciplines

Les chercheurs en psychologie ont souvent montré de l'intérêt pour l'éducation, et en particulier pour les technologies éducatives. Plusieurs chercheurs les ont considérées comme un support permettant l'application ou la mise à l'épreuve de leur théorie sur l'apprentissage. Ainsi, et déjà avant le développement de l'informatique, les machines à enseigner, développées par Skinner représentaient la mise en application directe de la théorie behavioriste de l'apprentissage selon laquelle les connaissances sont acquises par renforcement (Skinner, 1954). Newell et Simon ont introduit dans les années 70 Newell, et Simon, 1972) une nouvelle façon de concevoir et d'analyser la résolution de problème. Celle-ci est sous influence directe de l'avènement de l'informatique, comme celle de l'ordinateur, l'intelligence humaine est vue par les chercheurs comme la capacité à manipuler des symboles. La résolution de problème est représentée, dans ce nouveau cadre, comme le cheminement dans un espace de recherche. Les nœuds sont des situations du problème et les stratégies des règles de parcours. Ils développent alors un modèle capable de résoudre « n'importe quel problème » par une régression de but en sous-but qui porte le nom de General Problem Solver. Quelques années plus tard, ils écrivent sous le nom de Production System version G le premier système à base de règles de production (règles écrites sous la forme : si <condition> alors <action>). Cette approche est le parent commun de nombreux modèles cognitifs et d'intelligence artificielle (Neches, Langley, & Klahr, 1987). Parmi les héritiers, un modèle jouit d'une très grande popularité en psychologie cognitive et dans la modélisation de l'apprenant, c'est l'approche ACT qui change de nom au fil de son évolution (ACT*, ACTE) pour finalement s'appeler ACT-R (Adaptative Control of Thought — Rational) dont les numéros de versions augmentent au fil des ans (ACT-R 7 est à ce jour la dernière version). ACT-R est une architecture cognitive. Elle cherche donc à modéliser l'esprit humain d'une manière plus large qu'un modèle contextuel à un domaine. Dans cette architecture, les compétences sont modélisées par des règles de production activées par des connaissances déclaratives. L'architecture comprend de nombreux modules et mécanismes qui dépassent notre cadre d'analyse. ACT-R prévoit comment les règles de production sont initialisées, complétées et renforcées au cours de l'apprentissage du sujet en prenant en compte

certaines contraintes comme la mémoire de travail limitée. Pour John Anderson, les EIAH constituent un cadre idéal pour tester sa théorie de l'apprentissage. Ainsi, comme nous l'avons déjà pointé en introduction du manuscrit, Anderson et Ohlsson ont pu manifester dans le passé un enthousiasme sur le devenir symbiotique des EIAH et de la psychologie des apprentissages (Ohlsson, 1991, J. Anderson, 1984 cité dans Wenger, 1987). En effet, les environnements d'apprentissages ont de nombreux attraits. Ils permettent de bénéficier de données plus importantes que celles produites dans des conditions expérimentales classiques. Cela est d'autant plus vrai si l'environnement d'apprentissage dépasse le stade du prototype pour devenir un outil répandu dans les classes. Le deuxième attrait est lié à la condition d'expérimentation. Contrairement par exemple à ce que pourrait donner une étude passant par des livrets papier à compléter, il devient possible d'accéder à un certain nombre de variables comme le temps de réponse et les hésitations et le changement de paramètres pour générer des conditions d'expérimentations différentes. L'aspect dynamique des EIAH est à la fois intéressant d'un point de vue de recherche, mais aussi applicatif dans la mesure où il permet d'individualiser les apprentissages. De Skinner à Anderson, les technologies éducatives visent la personnalisation des apprentissages. La psychologie ne fournit pas seulement un modèle de l'élève, mais suggère un paradigme décrivant le moment et la forme des interactions humain-système. Le principe de certaines machines de Skinner était de libérer une sucrerie par l'emploi d'une trappe si l'élève répondait correctement à une question. Ce système ne nécessite pas la présence d'un professeur pour garantir son bon fonctionnement. Dans les tuteurs cognitifs construits sous l'approche d'Anderson, la résolution de l'élève est étudiée pas à pas pour la comparer avec l'ensemble des étapes qui permettent effectivement de résoudre le problème tel qu'un système expert à base de règle de production le permet. Cette décomposition est justifiée par la théorie du chercheur selon laquelle les connaissances peuvent être décrites à un niveau de granularité fin. Nous notons également un apport de la psychologie du développement dans le domaine des environnements d'apprentissage. Avec Piaget, la connaissance n'est plus conçue comme des informations transmises du professeur à l'élève (ou du système vers l'élève), mais comme une construction personnelle de l'élève, bâtie en interagissant avec son milieu. C'est l'approche constructiviste. Certains environnements sont conçus selon l'idée non pas de tutorer mais de fournir un environnement riche dans lequel l'enfant peut avoir un haut niveau d'interaction avec le système et observer le résultat de ses actions. Les approches dites néo-constructivistes ont aussi influencé les

environnements d'apprentissage, autant, si ce n'est plus, que l'approche originale de Piaget. Ainsi, Vygotski est souvent mentionné pour justifier l'importance des pairs dans l'apprentissage ainsi que du tutorat avec l'idée générale de fournir « un échafaudage » (scaffolding) pour aider l'enfant à résoudre le problème. Un autre concept populaire de Vygotsky, lié au concept d'échafaudage, est la zone proximale de développement. Entre les activités que l'enfant peut réaliser seul et celles qui lui sont inaccessibles se situe la zone de développement proximale. Elle définit les connaissances que l'enfant peut acquérir en présence d'une aide. Ainsi, l'importance du tutorat est soulignée chez Vygotski, ce qui le distingue de Piaget. Les approches basées sur les tuteurs intelligents se réclament régulièrement de Vygotski et les approches plus fondamentalement constructivistes comme les micro-mondes se réclament de Piaget.

La modélisation de l'apprenant dans les environnements d'apprentissage est-elle donc fondamentalement cognitive ? Cette conception idéaliste se heurte à un certain nombre d'écueils. Le premier, fondamental, est que l'environnement d'apprentissage a pour but de favoriser l'obtention de nouvelles connaissances par l'enfant. Modéliser l'apprenant n'est qu'un moyen pour parvenir à cette fin. Ainsi, certains auteurs dont Self, défendent l'idée que la modélisation de l'apprenant (Student Modeling) ne doit pas forcément être fondée psychologiquement. Par exemple, dans *Formal approaches to student modelling* (1994), Self oppose l'utilité des modèles à leur fidélité. Deux passages sont particulièrement révélateurs :

> *Student modelling may be "unabashedly psychological"(Clancey, 1986) but, as stated earlier, the primary aim is computational utility, not cognitive validity.(p. 19)*

Plus loin (p. 45):

> *Of course, the techniques and theories considered are justified to some extent by cognitive concerns but they can be developed and analysed independently of their cognitive content - just as a computational linguist may analyse grammars without commitment to their content or any view of human language use: an analogy which led to the coining of the term 'computational mathetics' for the kind of study presented here (Self, 1991b).*

Par ailleurs, Self (1990) lors de la conception d'un environnement d'apprentissage, suggère de ne pas intégrer dans le système de diagnostic des éléments inutilisés par la

suite (« *don't diagnose what you can't treat* »). De ce fait, il est assez fréquent dans les environnements d'apprentissage de limiter la modélisation au niveau de la réussite ou de l'échec d'une tâche. La position de Self est relativement consensuelle dans le domaine, nous pouvons par exemple lire dans la synthèse de Ragnemalm (1995, p. 4)

> *Knowing that the student doesn't know something is irrelevant if that knowledge cannot be used. So, presumably, the level of detail necessary in a student model corresponds to the remedial knowledge we are willing to put into the system.*

En d'autres termes, si une tâche est liée à des concepts appris, alors un système permettant de modéliser les connaissances de l'apprenant sans rentrer dans les détails de ses productions est concevable. Sous cet angle, la qualité d'un modèle est mesurée par le taux de bonnes prédictions concernant la réussite ou l'échec. Toutefois, cette conception du diagnostic dans les EIAH connaît des variations, Woolf (2009, p. 53) suggère par exemple qu'au contraire le diagnostic peut dépasser le cadre de l'interaction directe pour être utilisé et/ou analysé plus tard. Les prises de position sur le caractère cognitif du domaine et les possibilités de réciprocités imaginées citées au-dessus datent de deux à trois décennies. La question se pose de l'actualité de cette relation. Est-ce la position de Self (recherche d'efficacité) ou celle de Anderson et Ohlsson (tests de théories) qui a prévalu ? Concernant le lien direct (psychologie vers EIAH), depuis BUGGY, relativement peu de travaux cherchent à utiliser le cadre des environnements d'apprentissage comme l'occasion de tester des théories cognitives. Ohlsson maintient sa position sur la force de la relation dans un article récent (Ohlsson, 2016, p. 458) en suggérant que de bonnes théories d'apprentissages génèrent de bons tuteurs :

> *Twenty-five years ago, the constraint-based specialization theory of learning from error led to CBM, a novel and useful approach to ITS design. The fact that CBM derived from a theory of learning suggests that a better learning theory might generate an even better tutoring technology (Ohlsson, 1991).*

Deux remarques pourraient cependant être faites à l'encontre de cette position. La première est que le succès des CBT (Constraint Based Tutors) tient à la simplicité de leur développement. Ses promoteurs mettent en avant le fait que la tâche consistant à modéliser finement l'apprenant est évitée (Mitrovic, Koedinger, & Brent Martin, 2003). Ils se basent sur une théorie de l'apprentissage, certes, mais se caractérisent et se sont

posés en rupture de la modélisation plus précise de l'apprenant pour des raisons d'économie. La deuxième remarque est qu'il est difficile d'attribuer le succès CBT au modèle de Ohlsson. Les CBT montrent qu'un apprenant tire profit de ces erreurs, mais la théorie détaillant les processus de l'apprentissage par contraintes (Ohlsson & Rees, 1990) n'a, à notre connaissance, pas été directement testée au travers des différents environnements. Si nous pouvons concéder que les EIAH, en implémentant avec succès différentes stratégies pédagogiques, peuvent augmenter la plausibilité des théories dont elles dérivent, le niveau de preuve reste faible. Dans le cas des tuteurs cognitifs, une distance s'est créée entre l'architecture cognitive développée (ACT-R 7) et les tuteurs mis en place dans les EIAH : l'architecture cognitive a évolué, mais les tuteurs cognitifs restent à une version simplifiée de l'architecture (Dubois, Nkambou, Quintal, & Savard, 2010). Toutefois, en ce qui concerne les tuteurs cognitifs, il est notable que le niveau de précision des modèles en jeu permet aux chercheurs d'améliorer les connaissances sur les sources de difficultés des apprenants ainsi que leurs stratégies de résolutions (Ritter, Anderson, Koedinger, & Corbett, 2007). Il est donc possible de valider, au moins partiellement, le sentiment, presque prémonitoire, de Wenger (1987, p. 393) selon lequel la période d'enthousiasme pour les modèles d'apprenant précis touche à sa fin[10] pour des raisons d'efficacité. Les conclusions d'un état de l'art s'intéressant aux approches en modélisation de l'apprenant des dix dernières années (2003-2013) concluent dans ce sens. L'Overlay (recouvrement) est l'approche la plus utilisée, complémentée par des méthodes numériques pour la gestion de l'incertitude dans le diagnostic (Chrysafiadi & Virvou, 2013). Enfin, il est important de noter que le constructivisme de Piaget et le socioconstructivisme de Vygotski ont principalement influencé les environnements d'apprentissage de manière pédagogique seulement. Les approches constructivistes, bien que plus élaborées d'un point de vue cognitif que de simples recommandations éducatives, ne semblent pas avoir été très utilisées à un niveau de formalisation suffisamment avancé dans la modélisation de l'apprenant pour

[10] Citation exacte : *« In fact, after a period of excitement about student models, some researchers are expressing reservations about the assumption that more is better »*

pouvoir générer des diagnostics cognitifs[11]. Ohlsson (1988) explique cette absence par le fait que la tradition piagétienne a fourni des exemples de diagnostics, mais pas de méthode pour les réaliser.

## 2.3 Cas d'études

Pour exemplifier cet état de l'art, nous décrivons cinq systèmes. Compte tenu des objectifs de notre thèse, nous nous centrons sur la production de diagnostic cognitif en nous basant sur les techniques de modélisation du comportement qui se réclament d'un fondement cognitif. Ces outils sont étudiés de manière critique plus loin dans le document. Nous nous appuierons par la suite sur ces systèmes pour les étudier suivant leurs aspects cognitifs et statistiques pour former notre problématique.

### 2.3.1 Buggy

L'approche la plus connue dans l'emploi de bugs, responsable même de l'introduction de ce terme dans la littérature en modélisation de l'apprenant, est certainement le système BUGGY (Brown & Burton, 1978) devenu par la suite DEBUGGY (Burton, 1982). Dans ses premiers développements, l'application est conçue sous la forme d'un jeu destiné au professeur pour le former et le sensibiliser à la découverte de « *bugs* » dans les soustractions et additions posées par les élèves. Le répertoire de bugs a été initialement conçu à partir d'un modèle expert décrivant la suite d'opérations nécessaires à la résolution d'une opération comme un réseau de procédures. Ce modèle expert a été dégradé, non pas en enlevant des procédures (comme l'aurait fait un système de type overlay), mais en en remplaçant manuellement certaines par une version fautive. Le jeu BUGGY était conçu de la manière suivante : le système montre une opération erronée à l'utilisateur comme résultant d'un élève « *bugué* » qu'il s'agit de diagnostiquer en découvrant le bug. De multiples instances sont alors proposées

[11] L'approche cKc (Nicolas Balacheff, 1995) en proposant une formalisation de l'apprenant sous l'angle piagétien, représente toutefois un pas dans cette direction. L'utilisation du concept de théorème en acte proposé par Vergnaud (1982) dans le projet Aplusix (H. Chaachoua, Nicaud, & Bittar, 2005; Nicaud, Chaachoua, & Bittar, 2006) peut aussi compter comme exception.

jusqu'à ce que l'utilisateur découvre le bug qui génère ces erreurs. Outre ce jeu, BUGGY est doté d'un système de diagnostic recherchant, à partir de réponses d'élèves, les bugs ou combinaison de bugs présents dans le catalogue qui peuvent expliquer leurs erreurs. Le but de ce système est alors de détecter quel assemblage de règles erronées (appelés bugs) peut expliquer les erreurs des sujets dans des tâches de soustraction en colonne. Le caractère bruité des réponses (caractère instable des bugs) est pris en compte dans le diagnostic réalisé. DEBUGGY permet de construire dynamiquement et progressivement cet assemblage de bugs, limitant ainsi la complexité de l'algorithme. IDEBUGGY a ensuite été élaboré pour proposer des problèmes de manière dynamique, réduisant progressivement l'espace des diagnostics possibles. Par le biais de ces données, les chercheurs enrichissent leur base de bugs de manière itérative. Sur la base de ce travail ayant produit à la fois une importante base de bugs et un système de diagnostic automatique, des travaux ont eu lieu pour affiner le mécanisme de génération de bugs. VanLehn élabore une théorie nommée *« Repair Theory »* (Brown & VanLehn, 1980; VanLehn, 1990) proposant une explication de l'apparition des bugs. Cette théorie formalisée dans un programme informatique nommé Sierra, permet, avec une intervention humaine minimale, de générer des bugs à partir du modèle expert. Nous parlons alors d'une théorie « générative »[12]. Lorsque le sujet est en impasse, car les procédures qu'il connaît sont insuffisantes (par oubli ou non-connaissance des procédures expertes), il va effectuer une opération de réparation en employant une nouvelle procédure pour combler le vide provoqué par l'impasse. Ces opérations de réparation sont de quatre types et sont indépendantes de l'impasse. Cette hypothèse forte d'indépendance, permet, par la suppression-réparation de procédures du système expert de générer les bugs. DEBUGGY est ensuite utilisé pour compter, parmi les bugs prédits, lesquels sont effectivement observés dans les données, tout en comptant le nombre de bugs déjà connus, mais non prédits par Sierra.

[12] Le terme « génératif » est ici à prendre au sens de « grammaire générative » proposée par Chomsky dont la similarité d'approche est revendiquée dans la « repair theory ».

BUGGY est issu d'un travail de modélisation, dans lequel la validité psychologique de la modélisation par réseau de procédures « *procedural network* ». Brown et Burton l'indiquent dans une note en bas de page commençant par *« as a historical footnote* » BUGGY était pensé comme un instrument de recherche, pas un outil pédagogique. Ce n'est que plus tard que les auteurs ont réalisé que le système pouvait être d'intérêt pédagogique. Toutefois, il est possible de s'interroger sur la validité psychologique des bugs proposés dans BUGGY et Sierra. S'il semble que les bugs proposés dans les deux systèmes semblent effectivement en accord avec les données obtenues, la question de l'origine des bugs reste assez floue. Par ailleurs, le domaine est choisi pour être très orienté sur les connaissances procédurales. VanLehn le précise explicitement dans son livre (VanLehn, 1990, p. 12):

> *The main advantage of arithmetic procedures, from a methodological point of view, is that they are virtually meaningless to most students. They seem as isolated from commonsense intuitions as the nonsense syllables used in the Ebbinghaus paradigm for studying verbal learning*

Ce parti pris consistant à penser les procédures arithmétiques comme dénuées de sens permet de se concentrer sur les mécanismes de générations de bugs sans considérer la représentation psychologique des opérations. Or, même dans ce domaine jugé très procédural, les aspects sémantiques peuvent avoir du poids. Sander (2001) montre que cet axiome pris par VanLehn peut être remis en question. En effet, il propose que la soustraction soit basée sur deux analogies : ôter des éléments d'une collection et calculer la distance entre deux éléments. Ces représentations peuvent engendrer des bugs, très présents dans les données, et non générés par Sierra comme « 0-N=0 » (soustraction comme un retrait) ou « 0-N=N » (soustraction comme une distance). L'approche purement procédurale proposée par Sierra et BUGGY ne semble donc pas applicable à des problèmes beaucoup plus sémantiques comme les PAEV. Si nous pouvons reprocher la non-modélisation de ces aspects psychologiques, nous devons à la fois souligner l'ambition et l'importance de la démarche consistant à construire des modèles cognitifs générant des erreurs et des bugs pour les employer dans des démarches diagnostiques. Ces travaux ont maintenant plus de 30 ans, et pourtant, ils représentent des efforts en modélisation cognitive dont il est difficile de trouver des héritiers dans les approches modernes en modélisation de l'apprenant.

### 2.3.2 Emploi des règles de production

Les systèmes à base de règles de production sont des modèles embarquant des connaissances sous la forme de règles du type si <condition> alors <action>. Ces dernières ont été utilisées dans les travaux princeps en modélisation cognitive par Newell et Simon et le sont encore dans certaines approches héritières. Une des plus importantes, ACT-R avance la théorie que les compétences cognitives peuvent être formalisées par des règles de production. C'est une architecture cognitive qui a été adaptée pour les environnements d'apprentissage. Les tuteurs produits, nommés Cognitive Tutors ont marqué le domaine rapidement, et sont actuellement commercialisés à large échelle par Carnegie Learning Inc. Ils font partie de l'approche générale « Model Tracing ». En contraste avec le domaine des PAEV, le domaine de l'algèbre a connu beaucoup d'outils construits sur la base de règles de production, probablement du au fait que l'algèbre s'axe plus sur la manipulation de formules mathématiques que sur la représentation que le sujet peut se faire d'un texte. Certaines approches, comme Early Algebra Problem Solving (EAPS)(Koedinger & MacLaren, 2002), proposent l'emploi de règles de production pour expliquer la difficulté relative de différents problèmes d'algèbre. Les PAEV font partie des problèmes dont la résolution est simulée par EAPS. Dans ce cas, les règles de production couvrent à la fois la compréhension, la manipulation d'expression et le calcul. Quelques bugs dans le calcul sont aussi inclus. Des probabilités associées aux règles de production sont calculées permettant de construire un assemblage de règle dont les taux de succès se rapprochent de ceux généralement observés dans les différents types problèmes. EAPS permet donc de simuler l'« étudiant moyen ». SimStudent, projet faisant suite à EAPS, cherche quant à lui à simuler des individus particuliers (MacLaren & Koedinger, 2002) et permet le passage de la modélisation théorique aux modèles d'élèves simulables. Les PAEV sont dans EAPS considérés comme la version la plus simple des problèmes d'algèbre à énoncés verbaux et ne sont donc pas étudiés en profondeur.

Les compétences écrites dans EAPS sont variées, mais ne touchent que superficiellement à la compréhension de texte, en effet la compréhension est modélisée par des règles globales du type « comprendre le problème ». La modélisation dans EAPS se place donc juste en aval des compétences prévues d'être développées par l'apprentissage de la résolution de PAEV. Cette position des Cognitive Tutors pour l'algèbre est assez explicite dans l'article présentant EAPS et SimStudent (MacLaren & Koedinger, 2002, p. 359).

> *We chose not to model verbal comprehension in contrast to equation comprehension because beginning algebra students had little difficulty comprehending the word problems used*

La relation bidirectionnelle entre l'action et les représentations est difficile à représenter dans les systèmes de production. En effet, ces derniers ont pour postulat l'asymétrie entre les connaissances et les règles de production. La modélisation de l'impasse et des changements de représentation posent en général également question dans ce type de modèle. Ces modifications peuvent être modélisées comme l'activation tardive d'une règle, car les connaissances déclaratives qui lui sont nécessaires sont faiblement activées. En effet, les modèles à base de règles de production font l'économie du concept de représentation. Est-il alors possible de dire qu'ils ne sont pas adaptés pour modéliser la compréhension des problèmes ? Si cette critique est fréquente, ses défenseurs ne sont pas sans réponse. Anderson et Schunn (2000, p. 19) proposent une explication :

> *« In ACT-R, understanding a concept means nothing more or less than having a rich network of highly available declarative chunks and production rules that can be used to solve problems involving those concepts flexibly in many contexts. »*

Capturer le concept de représentation dans une architecture cognitive est une complication de taille. Dans ACT-R, le problème est donc « résolu » en décrivant les représentations comme des connaissances déclaratives existantes activées et des règles de production disponibles. La complexité de cette approche, et sa rigidité due à la difficulté d'implémenter un changement de représentation du problème ou le travail de compréhension face au problème explique peut-être pourquoi les tuteurs cognitifs (tuteurs basés sur ACT-R) sont pratiquement absents du champ de l'arithmétique en classe élémentaire. L'aspect procédural des systèmes à base de règles de production ne met pas l'accent sur le changement de représentation et les éventuels conflits dans la représentation du problème. Cette analyse est partagée par Reed dans un livre consacré aux PAEV (1999, p. 24):

> *It is not surprising that applications of production-system models to tutoring have focused on rule-based problem solving such as constructing a geometry proof or writing a short program in LISP. Anderson et al. (1995)*

> *have not attempted to model knowledge domains that are largely declarative.*

Il est important de concéder que les règles de productions, dans les travaux en psychologie de Riley et Greeno (1988), jouent un rôle important dans la modélisation des compétences. Le modèle construit sur la base d'une version de ACT, permet de simuler le comportement d'élèves en fonction de leurs capacités conceptuelles et permet de faire des prédictions sur la difficulté relative des PAEV. Toutefois, le problème soulevé précédemment reste présent.

### 2.3.3 Tuteurs basés sur des contraintes

Ces outils, connus sous le nom anglais « Constraint-Based Tutor » : CBT (Mitrovic et al., 2003; Ohlsson, 1994) reposent sur l'idée que l'élève apprend de ses erreurs. Ils se basent seulement sur un ensemble de contraintes qui permettent de détecter la non-validité d'une réponse. Si une contrainte n'est pas respectée, une rétroaction peut avoir lieu. Cette conception est d'un niveau de granularité un peu moins fin que celle d'Anderson, car elle porte sur la résolution de manière globale. Néanmoins, elle serait plus facile à mettre en place pour des résultats équivalents (Mitrovic, Mayo, Suraweera, & Brent Martin, 2001). En effet, les tuteurs utilisant un modèle complet de l'apprenant détiennent un rapport coût/bénéfice souvent peu avantageux freinant leur généralisation. Les contraintes sont formalisées par le couple {Cr, Cs} avec Cr représentant les cas où la contrainte est pertinente (r pour « relevant ») et Cs son critère de satisfaction. Nous reprenons l'exemple de contrainte fourni par Mitrović (1998, p. 3) :

> *If the problem is n1/d1 + n2/d2, the student's solution is n/d and n = n1 + n2, then it had better be the case that d1 = d2 = d.*

Ici le couple {Cr,Cs} est décliné de la manière suivante :

- Cr : Problème d'ajout de fraction pour laquelle la solution de l'élève a pour numérateur la somme des deux numérateurs
- Cs : Les deux dénominateurs des fractions à ajouter doivent être égaux.

À travers cet exemple, il est possible d'entrevoir l'usage des contraintes dans le cadre d'un tuteur. Lorsque cette contrainte est détectée comme violée, alors il est possible d'afficher un message d'erreur informatif, permettant à l'apprenant de comprendre en quoi sa réponse est erronée. Par exemple « *attention, pour additionner des fractions, il faut que les dénominateurs soient égaux* ». Les travaux, débutés par le développement

d'un tuteur pour l'apprentissage de la formation de requêtes SQL (Mitrović, 1998), ont été poursuivis par la construction d'outils auteurs du nom de WETAS et d'ASPIRE espérant généraliser l'emploi de tuteurs basés sur des contraintes dans l'éducation (Mitrovic, Brent Martin, & Suraweera, 2007).

Les tuteurs basés sur des contraintes ont une racine cognitive indéniable. En 1990, Ohlsson et Rees proposent et implémentent une théorie de l'apprentissage par l'erreur en se basant sur le concept de contraintes (Ohlsson & Rees, 1990). Plus tard, il est suggéré que les contraintes peuvent fournir un support efficace au tutorat dans les environnements d'apprentissage (Ohlsson, 1992, 1994). En 1998, avec le concours de Mitrovic, les premiers Constraint Based Tutor (CBT) sont implémentés. La théorie proposée par Ohlsson est une alternative aux systèmes basés sur les règles de production dans laquelle les connaissances procédurales sont déduites des connaissances déclaratives [13]. Au contraire, le mécanisme d'apprentissage de règles à partir de connaissances procédurales indiquant la présence d'erreur dans la solution semble plus simple, donc plus plausible. Les contraintes dans les CBT représentent alors une base plus grande de connaissances déclaratives que le sujet pourrait posséder. Le but premier de ces tuteurs n'est pas de faire apprendre des contraintes comme de nouvelles connaissances déclaratives, mais d'utiliser ces contraintes dans une stratégie d'apprentissage par correction d'erreur. L'intérêt des contraintes, d'un point de vue cognitif, est de donner une place plus importante aux aspects conceptuels dans l'apprentissage, notamment dans les mathématiques (Ohlsson, 2016). De ce point de vue, l'approche est importante pour notre recherche dans le domaine des PAEV. Toutefois, les contraintes, dans ce cadre, se limitent à définir « ce qui est vrai ». Elles ne sont pas utilisées, par exemple, pour représenter des conceptions erronées de l'apprenant (Brent Martin, Suraweera, & Mitrovic, 2007, p. 2) :

[13] Ohlsson (2016) explique à titre anecdotique que le jour où il a décidé de dériver les règles permettant de résoudre une soustraction posée à partir des règles d'algèbre, il a obtenu 25 pages de formules mathématiques et a été convaincu à partir de ce jour que ce mécanisme était peu plausible.

> *Constraint-based modeling uses abstraction to avoid the need to model students' misconceptions*

La non-modélisation des conceptions erronées est probablement pertinente dans le cadre d'un tuteur, mais est plutôt problématique si nous désirons construire un diagnostic précis du sujet. Certains travaux proposent de rechercher ces conceptions erronées par la fouille de donnée recherchant des phénomènes de cooccurrences de viol des contraintes dans les CBT (Elmadani, Mathews, & Mitrovic, 2012). D'autres chercheurs proposent d'améliorer la capacité diagnostique des CBT par un système de poids et de probabilités (Le & Pinkwart, 2011). Toutefois, ces approches restent limitées du fait que les CBT travaillent avec les contraintes du domaine, ce qui le différencie peu, dans la précision du diagnostic, des modèles de type Overlay (recouvrement). En conclusion, si les aspects cognitifs des CBT sont importants, la profondeur des diagnostics est faible du fait que le système se concentre sur la détection de déviation à la normalité.

### 2.3.4 Modèle des contraintes

Dans les modèles symboliques dits computationnels, le diagnostic est réalisé par la recherche de composantes symboliques élémentaires (des bugs en l'occurrence). Le modèle des contraintes est un de ces systèmes. Dans ce modèle, la brique élémentaire modélisant le comportement est le concept de contrainte. Il utilise le concept d'espace de recherche introduit par Newell et Simon (1971). Il se décompose en nœuds qui contiennent tous ses états possibles. Un graphe est construit en considérant toutes les transitions existantes. L'espace de recherche permet donc de formaliser le comportement de résolution comme un déplacement dans ce graphe. Les premiers systèmes ont proposé de formuler le comportement comme des règles de production : si <condition> alors <action> laissant, à chaque nœud, une transition unique possible. Le modèle des contraintes, lui, choisit de se baser sur une formalisation opposée : si <condition> alors Non <action>. D'un point de vue formel, c'est une généralisation des règles de production, puisqu'une règle de production peut être vue comme une contrainte qui n'autorise qu'une seule action.

Le modèle des contraintes a été employé dans la modélisation de la résolution du problème des tours de Hanoï (Richard, Poitrenaud, & Tijus, 1993). Dans ce modèle, les contraintes permettent à la fois de décrire les règles de déplacement des disques (tel que le sujet le perçoit), des heuristiques (ex. : ne pas déplacer deux fois le même disque), mais aussi des buts. Les règles de déplacement des disques sont donc des contraintes

comme les autres. En cas d'impasse, elles peuvent donc être relâchées. Dans ce cas, certains sujets viennent à produire des actions illégales lorsque peu de choix s'offrent à eux. Ainsi, les représentations du sujet viennent à changer lorsque l'ensemble des actions possibles est épuisé. Cette spécificité du modèle des contraintes permet de formaliser le phénomène de changement de représentation. En d'autres termes, le concept de représentation n'est pas un acquis stable, mais une composante de la résolution de problème sujette à évoluer. Dans la modélisation de la résolution des tours de Hanoï, les sujets n'étaient pas diagnostiqués par le modèle. Ce dernier simulait pas à pas le comportement de résolution général par le biais de contraintes en prédisant ses actions du sujet au fil de l'eau. Dans d'autres applications, le modèle des contraintes a été utilisé pour diagnostiquer les différents apprenants (Richard, Pastré, & Parage, 2009). Il s'agit alors de décrire des contraintes portant sur les stratégies de résolution, mais aussi extraites des connaissances expertes ou erronées dans le cadre d'une tâche professionnelle complexe. Elle consistait au réglage de presse à injecter. Des analyses préalables ont permis d'obtenir une connaissance des actions réalisables dans le cadre de cette tâche ainsi que les connaissances expertes (Pastré, 1994 cité dans Richard et al., 2009). Après formalisation de l'espace de recherche décrivant les différentes opérations et le graphe des états du problème, les connaissances expertes, leurs formes dégradées, ainsi que les différentes stratégies de réglages ont été traduits par des contraintes cognitives. Par un mécanisme de sélection, les contraintes les plus explicatives des protocoles des apprenants sont choisies et s'agrègent pour former le diagnostic de chaque sujet. Le concept de contrainte est très utilisé en psychologie cognitive dans les problèmes qui nécessitent un changement de représentation (Choi & Ohlsson, 2010; Kershaw, Flynn, & Gordon, 2013; Knoblich, Ohlsson, Haider, & Rhenius, 1999; Öllinger, Jones, Faber, & Knoblich, 2013; Patrick & Ahmed, 2014). Les cas paradigmatiques incarnant ce phénomène sont les problèmes à « Insights », c'est à dire des problèmes pour lesquels le sujet s'aperçoit subitement de la solution d'un problème ou de la marche à suivre pour l'obtenir. La Figure 1 représente un de ces problèmes. La consigne est de relier ces 9 points en formant quatre segments de droite sans lever le crayon. À gauche est représenté le problème, à droite la solution de ce problème. Le problème est difficile car les sujets se donnent la contrainte de ne pas déborder. Or il est nécessaire de « sortir du carré » pour trouver la solution.

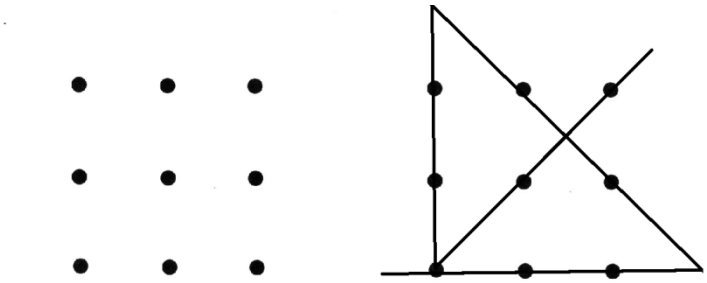


Figure 1. Figure et solution du problème des 9 points (Maier, 1930).

Le modèle de contraintes (Richard et al., 2009, 1993) se propose de modéliser le comportement en mettant au centre de son formalisme le concept de contraintes. Le formalisme peut être vu comme une extension des règles de production (si <condition> alors **NON** <action>). Ainsi, si la notion d'impasse dans ACT-R est l'idée d'un *« pas assez »*, dans le modèle des contraintes il s'agit d'un « *trop* ». Il ne s'agit pas d'activer une nouvelle règle, mais de relâcher une contrainte. Le modèle des contraintes est une théorie négative de l'action dans le sens où les représentations du problème sont décrites comme des limitations des actions possibles. La notion de conflit dans ACT-R survient lors de l'activation concourante de plusieurs règles à la différence du modèle des contraintes, où le conflit se produit lorsque le problème est surcontraint. En contrepartie, il offre la possibilité de modélisation des représentations et leurs aspects conflictuels à la différence des systèmes à base de règles. Les contraintes permettent ainsi d'expliquer simplement la relation symétrique entre les possibilités d'actions et les représentations. Si ces dernières se constituent face à notre représentation des problèmes, en retour nos possibilités d'actions la modifient. L'aspect économique d'un point de vue cognitif des contraintes est un point important. Dans les règles de production, s'il s'agit de construire un système évitant de saisir des objets brûlants, cette condition doit être inscrite pour chaque règle de production concernant les actions passant par le toucher d'objet, ce qui est coûteux.

Le changement de formalisme entre les systèmes à base de règle de production permet aussi un revirement dans la méthode d'analyse du comportement. À chaque nouvelle contrainte implémentée dans le modèle, le champ des actions prédites se réduit, ce qui permet de modéliser le sujet de manière itérative. Par accumulation des restrictions, les contraintes se cumulent facilement ce qui permet de modéliser un comportement avec des composantes hétérogènes de la résolution de problèmes : buts, stratégies, représentations, connaissances, et mêmes règles de production (qui peuvent être représentées comme des contraintes n'autorisant qu'une seule réponse). Il est alors

possible de concilier, dans un même modèle, des approches théoriques relativement différentes (Richard et al., 1993). Cette qualité est très importante au vu de la variété des approches théoriques de la résolution de PAEV. Nous favoriserons donc le concept de contrainte au concept de règle dans nos travaux de modélisation cognitive (chapitre 1 et 3 des contributions). Bien que le concept de contrainte ait une place fondamentale dans l'approche du modèle des contraintes et des tuteurs basés sur des contraintes, l'emploi diffère radicalement. Dans les tuteurs, les contraintes décrivent le domaine. L'aspect cognitif de l'approche de Mitrovic et Ohlsson provient d'une théorie sur l'apprentissage à partir des erreurs et, dans le cas général, les contraintes ne constituent pas le modèle de l'apprenant, au contraire le modèle des contraintes est plutôt un modèle recherchant la simulation du comportement en résolution de problème. Cette différence permet aussi d'expliquer la différence de notation entre les deux approches (couple {Cr,Cs} pour les CBT et si <condition> alors non <action> pour le modèle des contraintes). Dans l'approche d'Ohlsson, les contraintes représentent les connaissances déclaratives alors que le modèle des contraintes utilise une définition étendue en représentant les représentations et les stratégies du sujet, rendant le modèle exécutable, à l'opposé de la définition de Ohlsson (Ohlsson, 1994, p. 184). Il est toutefois possible, dans la conception du modèle des contraintes, d'introduire des contraintes décrivant le domaine. Cette transversalité du concept de contrainte est extrêmement encourageante du point de vue de la modélisation de l'apprenant dans les environnements d'apprentissage.

### 2.3.5 L'Item Response Theory et les conceptions erronées

Les quatre exemples précédents font partie de la famille des modèles symboliques. Comme nous l'avons introduit plus tôt, les approches numériques sont maintenant très utilisées dans le champ de la modélisation de l'apprenant. D'origine psychométrique, la théorie des traits latents, appelés maintenant l'Item Response Theory couvre une famille de techniques de modélisation majeure du domaine. Dans sa version la plus classique, la probabilité qu'un Individu I résolve le Problème P est modélisée par une régression logistique utilisant un ou plusieurs paramètres propres à l'individu (des traits latents) et un ou plusieurs paramètres qui dépendent du problème. Dans le cas d'une fonction logistique à 3 paramètres, ces derniers sont propres à l'exercice, lorsqu'ils sont complémentés avec un paramètre quantifiant la capacité de l'apprenant ($\theta$), la probabilité de réussir l'exercice p est donnée par l'équation suivante :

$$p = c + (1 - c)\frac{e^{a\theta - b}}{1 + e^{a\theta - b}}$$

La Figure 2 présente graphiquement cette équation en prenant c=0,25 ; a=1 ; b=0

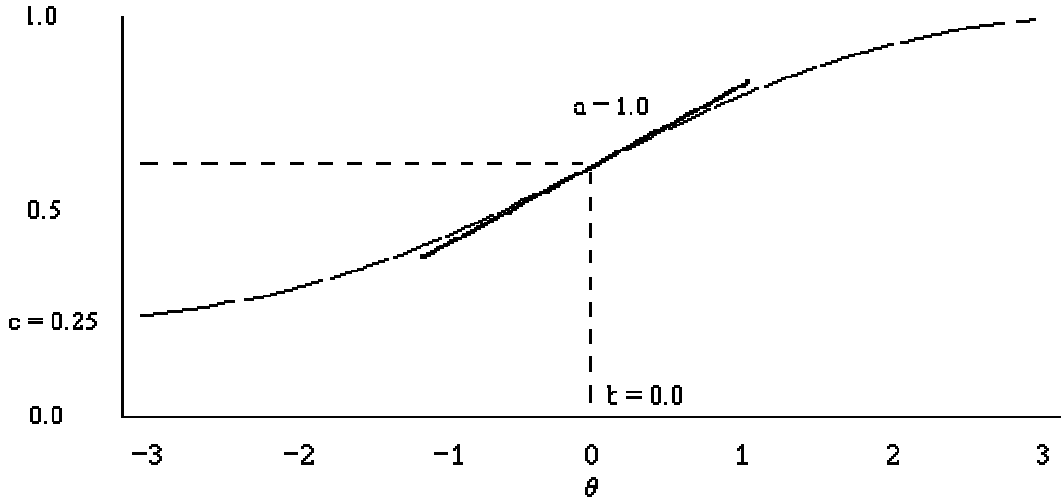



Figure 2. Figure présentant une fonction logit à trois paramètres[14]

Cette approche a eu un fort succès dans la modélisation de l'apprenant du fait de son aspect pratique. Brian Junker (1999, p. 21), dans sa revue sur les méthodes statistiques pour les systèmes de diagnostic, résume la qualité des approches psychométriques par le calcul systématique de mesures attestant de la validité (validity) et de la fiabilité (reliability) des modèles testés.

> *Two touchstones of traditional psychometrics are validity, the extent to which we are measuring what we intend to measure, and reliability, the extent to which observable responses are well-determined by underlying*

[14] **Source : "3PL IRF" by Iulus_Ascanius. Licensed under Public Domain via Commons — https://commons.wikimedia.org/wiki/File:3PL_IRF.png#/media/File:3PL_IRF**

> *student variables (attributes or proficiencies), and conversely the extent to which inferences about these latent student variables would be stable if the assessment were repeated.*

Depuis les travaux de Tatsuoka (1983, 1990), la psychométrie trouve une place importante dans l'établissement de diagnostic cognitif par l'emploi de Q-matrices. Elles sont généralement construites par des experts du domaine et établissent un lien entre des composantes cognitives et la capacité de résolution d'un exercice. Si « 1 » est à l'intersection du problème et de la composante cognitive, alors elle est nécessaire à la résolution du problème. Le Tableau 1 est un exemple de Q-matrice : C1 représente la capacité à calculer la soustraction d'un nombre négatif; C2 la capacité de soustraire un nombre à un nombre plus petit; C3 la capacité d'additionner un nombre négatif à un nombre positif. Lorsqu'une cellule a un 1, la compétence est nécessaire à la résolution de l'exercice.

Tableau 1. Exemple réduit de Q-matrice.

| item | C1 | C2 | C3 |
|---|---|---|---|
| 3-(-4) | 1 | 0 | 0 |
| -2+8 | 0 | 0 | 1 |
| 1-5 | 0 | 1 | 0 |
| -4+8 | 0 | 0 | 1 |

Or, Tatsuoka, en plus d'ouvrir la voie aux Q-matrices, a justement proposé une méthodologie basée sur le diagnostic de règles. Sa méthode, nommée Rule Space Method (RSM) consiste à classer les apprenants dans des groupes prédéterminés par de possibles connaissances erronées. Son approche est hybride. D'une part, elle utilise l'outillage statistique de l'IRT, qu'elle complète avec une autre mesure continue, $\zeta$ qui est orthogonale à $\theta$[15] et mesure de l'« *atypicalité* » des performances. La simulation d'élèves utilisant des règles erronées permet d'associer l'emploi de ces règles à un point

[15] Dans le cadre des régressions logistiques les plus simples, $\theta$ représente le trait latent représentant la capacité générale de l'apprenant.

dans l'espace ($\zeta$, $\theta$). Sa méthode est graphique, si l'apprenant est dans une zone proche de ce point, alors il est diagnostiqué comme ayant la conception erronée liée à ce point.

L'approche utilisant les Q-matrices, par exemple, se revendique cognitive[16] puisque cette matrice décrit une relation entre les connaissances du sujet et les exercices qu'il peut réussir. Cette matrice est en effet généralement construite par une analyse de la littérature ou par des experts du domaine[17]. L'approche RSM de Tatsuoka que nous avons vue précédemment consistant à diagnostiquer des règles a été jusqu'à présent peu suivie dans les EIAH, au profit d'une utilisation des Q-matrices pour les **compétences uniquement**, en témoigne la première ligne d'un article sur le domaine (DeCarlo, 2011, p. 8) :

> *Cognitive diagnostic models (CDMs) attempt to uncover latent skills or attributes that examinees must possess in order to answer test items correctly.*

Les termes CDM « *Cognitive Diagnostic Model* » et CDA « *Cognitive Diagnostic Assessment* »[18] sont aujourd'hui le plus souvent accompagnés des Q-matrices. La récente revue de la littérature de DiBello, Roussos, & Stout (2006) peut en témoigner : sous le nom de « *Cognitive Diagnostic Assessment* », plus d'une douzaine de modèles présentant des méthodes statistiques psychométriques variées basées sur les Q-matrices sont comparés. Cet emploi du terme cognition peut surprendre lorsqu'elle est observée du point de vue de la psychologie cognitive. Cette différence de connotation est toutefois soulignée par certains chercheurs du domaine (Rupp & Templin, 2008, p. 221):

---

[16] Tatsuoka (2009, p. 50) : *« As Chipman, Nichols, and Brennan (1985) stated, RSM is a hypothesis testing of cognitive models. »*

[17] Il convient de noter ici que des travaux récents en Educational Data Mining visent à aider à la conception de ce type de matrice à partir des données (Desmarais, Beheshti, & Naceur, 2012; Desmarais & Naceur, 2013)

[18] Les labels alternatifs portant sur ces méthodes sont en plus grand nombre et peuvent porter des connotations différentes (Rupp & Templin, 2008 pour une vue d'ensemble)

> *As discussed, for example, by Rupp (2007) and Mislevy (2007), the connotations of the word "cognition" in educational assessment differ from the core meaning of the word within the discipline of cognitive psychology*

La psychologie cognitive a toutefois nourri ce type d'approche psychométrique (DiBello, Stout, & Roussos, 1995). Cependant, du fait que notre thèse est motivée par la recherche de techniques de modélisation bénéficiant à la fois aux EIAH et à la recherche en psychologie, il convient de se poser la question sur l'influence des méthodes de diagnostic psychométriques pour la psychologie cognitive. Or, elle est considérée comme quasi inexistante par certains chercheurs du domaine (Rupp & Templin, 2008, p. 224) :

> *Moreover, while research in applied cognitive psychology informs research and practice in educational assessment, the opposite flow of information is effectively nonexistent.*

Pourtant, certains chercheurs ont insisté sur l'idée que la recherche en « educationnal assesment » devrait entretenir un lien plus fort avec la psychologie cognitive, mais que le formalisme utilisé est inadéquat. Lord's (1980), cité dans (J. Leighton & Gierl, 2007, p. 9) pointe, en faisant référence aux régressions logistiques, que la complexité de la résolution de problème ne peut pas être résumée par trois valeurs numériques. Il est remarquable que, tout comme le domaine des environnements d'apprentissage, des chercheurs aient manifesté un certain optimisme dans les années 80 (Snow & Lohman, 1989) sur la relation qui pouvait être entretenue avec la psychologie cognitive et que le constat, deux décennies plus tard, soit pour le moins mitigé (J. Leighton & Gierl, 2007, pp. 14–15).

> *As Snow and Lohman (1989) mentioned, the challenge for cognitive psychologists is to produce such theories, based on sound scientific studies, for use in educational measurement. But this challenge has not been met. Not because cognitive psychologists are uninterested in measurement, because indeed many are (e.g., J. R.Anderson, R. J. Sternberg), but because scholars focus first on investigating the research questions of their own discipline using their own methods and tools (Sternberg, 1984) before embarking on research questions in other disciplines.*

Cette remarque, à notre sens, est fondamentale et aurait pu être reprise presque mot pour mot pour décrire la relation avec la psychologie cognitive et les environnements

d'apprentissage. Or, comme les auteurs l'expliquent ailleurs, l'utilisation de régressions logistiques est peu compatible avec la manière dont le comportement est modélisé en psychologie cognitive. La psychologie cognitive a tendance à s'intéresser aux processus alors que la psychométrie s'intéresse aux facteurs ou traits latents. De ce fait, il est compréhensible après une analyse extensive de la littérature en modélisation cognitive, Leighton et Gierl concluent qu'il n'y a pas de modèles cognitifs disponibles clef en main pour la psychométrie (J. P. Leighton & Gierl, 2011, p. 200).

### 2.3.6 La Rule Space Method et la psychologie cognitive

Les approches psychométriques représentent l'exact opposé des approches symboliques. Leur bagage statistique est bien établi, mais ils peinent à représenter des aspects de la cognition dépassant la problématique de la prédiction du succès à un problème donné. Dans ses formulations classiques, l'IRT ne permet pas de modéliser et diagnostiquer des conceptions erronées de l'apprenant. Toutefois, des alternatives, dont la Rule Space Method (RSM), ont été proposées pour enrichir le diagnostic par ce type d'indicateurs. Comme nous l'avons vu précédemment, Tatsuoka a proposé une étape supplémentaire qui consiste à **diagnostiquer la présence de règle** par une méthode graphique. Néanmoins, la richesse des règles décrites est largement sous-exploitée. Alors que ces dernières pourraient produire des prédictions précises sur le type de réponses obtenues lorsqu'elles sont respectées, elles ne sont utilisées que par l'intermédiaire de la réussite ou l'échec des problèmes.

Prenons la règle suivante : « *Lorsqu'une opération engage deux nombres relatifs, une addition est faite et le signe du plus grand nombre est choisi* ». Une telle règle donnerait des réponses justes pour les réponses du type –14 + –5 = ? et fausses pour les réponses du type 3 + –5 = ?. Dans l'approche de Tatsuoka, une faute sur la question 3 + –5 = ? « contribue » à cette règle, quel que soit le type d'erreur. C'est une perte notable, car sur ce problème, si la réponse -8 fournit un support pour la règle en question, c'est aussi le cas pour les réponses 2 ou 8 qui ne sont pas vraiment en accord. Par ailleurs, cette stratégie rajoute une complexité sur une approche déjà lourde à mener.

# 3 ÉVALUATION DES DIAGNOSTICS COGNITIFS

La question de l'évaluation des diagnostics est centrale dans notre questionnement. La capacité d'évaluer statistiquement un diagnostic est cruciale s'il est désiré qu'il ait une utilité dans le cadre d'une recherche en psychologie cognitive pour valider les modèles cognitifs qui les produisent. L'enjeu est tout aussi important pour les EIAH. Si nous voulons défendre l'intérêt des systèmes de diagnostic d'un point de vue pédagogique, le diagnostic produit a tout intérêt d'être associé à une pertinence qui est quantifiable. Or, comme nous le verrons dans le cadre de cette partie, c'est un point faible des modèles symboliques dont le caractère déterministe des prédictions est souvent reproché par les promoteurs d'approches numériques du diagnostic cognitif (ex. : Tatsuoka, 1985). Si, compte tenu des parties précédentes, les modèles symboliques semblent adaptés pour décrire l'apprenant d'un point de vue fertile pour la psychologie cognitive, la question de l'évaluation de ce type de modèle est problématique.

## 3.1 Évaluation des diagnostics dans différents cas d'études

### 3.1.1 Méthode d'évaluation des diagnostics dans BUGGY

BUGGY et ses successeurs (Brown & Burton, 1978) constituent leurs diagnostics en cherchant le meilleur assemblage de règles simulant les réponses de l'apprenant. Du fait du nombre important de règles possibles (plus d'une centaine), un tel diagnostic ne serait pas possible si elles ne faisaient pas des prédictions précises sur le type d'erreurs qu'elles engendrent. Toutefois, en matière d'expérimentation humaine, les données sont

bruitées, identifier un bug nécessite de prendre en compte la possibilité qu'il ne soit pas respecté sur l'ensemble des problèmes.

Dans BUGGY, une règle est considérée comme :

- Stable si elle est en accord au moins 75 % du temps sur les problèmes dans lesquels une erreur est prédite.
- Présente mais instable si :
    - Ce seuil est entre 50 % et 75 % (avec la condition supplémentaire que deux erreurs au moins soient prédites avec succès).
    - Une autre possibilité pour qu'un bug appartienne à cette classe est que le bug prédise plus de la moitié des erreurs spécifiques de l'élève, et se trompe dans moins de 50 % des cas lorsqu'il prédit une erreur alors que l'élève a répondu correctement.

Une approche semblable à BUGGY est le module de diagnostic développé dans le cadre du projet Aplusix (H. Chaachoua, Nicaud, & Bittar, 2005; Nicaud, Chaachoua, & Bittar, 2006). Des règles portant sur les transformations d'équations algébriques, alors appelées théorèmes en acte (Vergnaud, 1982), sont identifiées. À la différence de BUGGY, les règles de transformations ne prédisent pas de réponse unique, mais peuvent être cohérentes avec un certain nombre de réponses. Tout comme BUGGY, un système relativement arbitraire de seuils est mis en place. Dans Aplusix (H. Chaachoua et al., 2005), une règle (ou son opposée faisant les prédictions contraires) peut être constitutive du diagnostic à plusieurs niveaux :

- Elle est considérée comme stable si elle est en accord avec 75 % des situations dans lesquelles elle peut être employée et si leur nombre est supérieur à 4.
- Dans le cas où le nombre de situations concernées par la règle est inférieur à 4, les données sont jugées insuffisantes pour pouvoir juger si la règle est respectée par l'apprenant.
- Le dernier cas correspond à n>3 et le seuil des 75 % n'est ni atteint par la règle ni par celle opposée, alors elle n'est pas retenue.

Ces systèmes de décisions sont un peu difficiles à suivre et peuvent être qualifiés d'arbitraires. L'utilisation de seuil est une source de critique notable des modèles non probabilistes, car la pertinence du système de décision est difficile à établir. Ces

systèmes sont cependant justifiables dans la mesure où il est nécessaire de fixer un ou plusieurs seuils pour décider si une règle intègre ou non le diagnostic final. Les concepteurs de BUGGY admettent le caractère arbitraire de ces seuils, mais expliquent dans une note en bas de page que ce système de classification a été confirmé puisqu'il donne les mêmes résultats qu'une classification établie à la main.

### 3.1.2 Méthode d'évaluation du diagnostic dans le Modèle des contraintes

Comme nous l'avons précisé précédemment, le modèle des contraintes a été utilisé pour générer des diagnostics dans le cadre d'une résolution de problème. Le modèle recherche le meilleur assemblage de contraintes, c'est-à-dire celui qui simule le mieux le comportement de chaque individu. L'ajustement du modèle est rapporté par ces deux valeurs : le pourcentage d'erreur (cas où l'action d'un individu est incompatible avec son diagnostic) et le pourcentage de cas où le diagnostic n'autorise qu'une et une seule option[19] (Richard et al., 2009). Il est souhaitable, pour que le modèle soit jugé bon, que :

- Le pourcentage d'erreur soit bas.
- Le nombre de réponses compatibles avec le diagnostic à chaque instant soit faible.

À chaque contrainte ajoutée dans le modèle général, par définition, le taux d'erreur ne peut que monter et le nombre de réponses compatibles (que nous appellerons taille de l'étau dans la suite) ne peut que diminuer. Comment alors arbitrer le choix d'ajout ou de suppression de contraintes dans le modèle général si une mesure s'améliore et une autre s'empire? La question des seuils pour sélectionner deux modèles concurrents, est complexe et se heurte au même problème d'arbitrarité que la sélection des règles dans BUGGY. Un autre problème, commun aux deux est la prise en compte du nombre de règles (ou de contraintes) présent dans le système de diagnostic. Si, avec une taille des données constante, le nombre de règles disponibles est trop important, alors le risque de faux positif augmente. En d'autres termes, il serait possible de trouver une adéquation

[19] Le diagnostic étant un ensemble de contraintes, il est naturel que plusieurs options restent compatibles avec celui-ci.

entre le système de diagnostic et des données complètement aléatoires avec suffisamment de règles.

### 3.1.3 Méthode d'évaluation de productions de bugs dans Sierra

Le programme Sierra n'est pas à proprement parler un outil permettant d'établir des diagnostics. Toutefois, les bugs produits dans Sierra sont ensuite utilisés dans les programmes (I)(DE)BUGGY donc les méthodes employées pour juger de la qualité des productions de Sierra est fondamentale pour la qualité des systèmes de diagnostic. VanLehn s'est inspiré de l'approche générative de Chomsky, et emprunte le concept de validité observationnelle (observational adequacy) comme mesure de la qualité des prédictions de Sierra (VanLehn, 1990, p. 167). Il s'agit d'évaluer le recoupement entre les bugs prédits et les bugs observés. Il est désirable que le nombre de bugs prédits non observés soit faible et que le ratio de bugs prédits et observés sur le nombre de bugs observés total soit grand. VanLehn précise que cette approche ne permet pas de quantifier la qualité du modèle dans l'absolu, mais qu'elle permet de comparer la qualité de deux modèles concurrents. Cependant, le fait d'avoir plusieurs valeurs pour quantifier la pertinence d'un modèle peut mener à des indécidables. Le problème est similaire à celui rencontré avec le modèle des contraintes que dire par exemple d'un modèle alternatif qui produirait plus de bugs observés, mais aussi plus de faux positif ?

### 3.1.4 Méthode d'évaluation du diagnostic dans les approches psychométriques

En se plaçant dans le paradigme des régressions logistiques, l'évaluation des modèles psychométriques est très bien établie et consensuelle. Toutefois, les approches psychométriques ont un nombre parfois important de degrés de liberté qui impose une grande quantité de données pour la validation de leurs modèles. L'approche RSM de Tatsuoka est reconnue comme économique, car le diagnostic des conceptions erronées est réalisé à l'issue d'une classification et n'augmente pas le nombre de paramètres libres des modèles. Cependant, nous avons précédemment suggéré que cette méthode de diagnostic pouvait manquer d'efficacité. Toujours sur la base d'une extension des Q-matrices, d'autres approches telles que le « Rule Space Method » de Tatsuoka consistant à rechercher les conceptions erronées dans les données ont été proposées. Des méthodes telles que le « Scaling Individuals and Classifying Misconceptions » (SICM) permettent de prendre en compte le type de réponse donnée par l'apprenant (pas seulement réussite/échec), mais elles peuvent conduire à une quantité de paramètres

libres très importante rendant l'approche très difficile à déployer dans la pratique, car nécessitant une quantité de données par sujet importante (Bradshaw & Templin, 2014). En effet, en plus des paramètres continus représentant les compétences, des paramètres discrets sont rajoutés au modèle pour modéliser les conceptions erronées[20]. Le problème des paramètres libres et de la taille des données est reconnu comme un défi dans la psychométrie pour le diagnostic de conceptions erronées. Le champ dispose de modèles économiques, mais pouvant manquer d'efficacité comme le RSM et des modèles plus lourds détenant bien plus de paramètres libres (Rupp & Templin, 2008).

### 3.1.5 ASPM et ses critiques

ASPM (Analysis of Symbolic Parameters Model) est un programme qui a contribué aux développements des successeurs de BUGGY (Simon, Polk, & Vanlehn, 1995). Son principe est de conserver les paramètres symboliques que constituent la présence ou l'absence de règles dans le diagnostic. ASPM, « résout » le problème des seuils arbitraires dans les modèles symboliques précédents. Le programme propose un algorithme permettant de sélectionner le meilleur assemblage de règles pour construire un diagnostic. De ce fait, la mesure à optimiser est la distance de Hamming (qui compte le nombre de prédictions erronées). Deux problèmes se posent avec cette mesure de qualité. ASPM est conçu pour des modèles à base de règles de production, or dans des modèles pour lesquels de multiples observations peuvent être en accord avec le profil de l'élève, comme les modèles à base de contraintes, la taille de l'étau, c'est-à-dire le nombre de prédictions alternatives laissées par l'agrégat de contraintes, doit être prise en compte. Le deuxième problème est que cette mesure ne tient pas compte du nombre de paramètres libres du modèle, en effet le modèle pourrait trouver le meilleur assemblage de règle sur des données complètement aléatoires sans « *s'en apercevoir* ».

[20] Il serait d'ailleurs possible, dans une perspective de modélisation cognitive de questionner le sens de la dissociation proposant des valeurs continues pour les compétences et des valeurs discrètes pour les conceptions erronées. Une telle dissociation semble avoir pour implicite que la « bonne et la mauvaise connaissance » sont de natures différentes et que ces différences de natures peuvent être expliquée par l'aspect discret ou numérique du paramètre qui les représente.

Les éditeurs du livre dans lequel la présentation d'ASPM a paru ne cachent pas leur avis plutôt mitigé. Ils présentent le chapitre d'une manière, certes élogieuse en premier abord, mais le qualifient ensuite de « plutôt extrême ». Ils disent regretter que les phénomènes stochastiques ne soient pas pris en compte et que la qualité d'un diagnostic doive être mesurée (Nichols, Chipman, & Brennan, 1995, p. 6).

> *This is a totally deterministic approach, and it obviously depends on having a very high-quality cognitive model. Of course. by identifying areas in which there are failures of fit, the method can be used to help refine a cognitive model. Still, a completely deterministic approach is rather extreme. As yet, ASPM does not have stochastic aspects that would provide for accidental slips in examinee performance or that would enable one to decide whether a diagnosis with less than perfect fit should be considered good enough. This is an important direction for future developments of ASPM.*

Les perspectives suggérées par les éditeurs dans les deux dernières phrases ne sont pas vraiment celles des auteurs du chapitre (qui sont plutôt d'ordre technique), ce qui rend ce passage encore plus intrigant. La tension entre les approches symboliques et numériques est donc palpable. Le projet ASPM, qualifié de « *rather extreme* », n'a pas été continué, et la recherche de publications citant les articles présentant ASPM n'a pas permis de trouver une approche s'en réclamant ou cherchant à l'étendre.

## 3.2 Le problème des degrés de liberté.

Avant de proposer une nouvelle méthode de conception de modèles cognitifs dans les EIAH, il est important de pouvoir fonder une mesure qui puisse établir leur pertinence. Or, la question de l'évaluation d'un diagnostic est indissociable de la question du nombre de degrés de liberté.

### 3.2.1 Problème général de la complexité des modèles

La question de la limitation des degrés de liberté est centrale dans la modélisation. Recourir aux modèles les plus simples possible est parfois représenté sous l'ancienne notion philosophique du rasoir d'Occam. La notion de minimum de complexité dans la modélisation peut avoir de multiples sens. Elle peut être vue comme un principe à suivre lors de la construction de modèles. Dans une vision poppérienne de la méthode scientifique, un modèle simple est plus facile à falsifier qu'un modèle étoffé

d'hypothèses auxiliaires, il est donc « plus scientifique ». Une autre raison, alternative, est de considérer qu'à précision égale, les modèles les plus simples ont plus de chances de réaliser de meilleures inductions que des modèles complexes.

Cette idée peut-être débattue d'un point de vue philosophique lorsque la complexité est prise dans son sens abstrait, en partie parce que la simplicité d'un modèle est un jugement que nous pouvons qualifier de subjectif. S'il s'agit de complexité en tant que nombre de paramètres libres, alors il existe généralement un consensus : si deux modèles expliquent aussi bien des données, alors celui qui en a le moins doit être favorisé. Au contraire, un modèle qui a beaucoup de degrés de liberté peut avoir une précision très forte, c'est un problème que l'on nomme risque d'overfitting, traduit parfois en français par « surajustement ». Par exemple, il est toujours possible, par le biais d'une interpolation polynomiale, de trouver une courbe qui passe exactement par N points d'un plan en utilisant un polynôme de degré N, c'est ce qu'illustre la Figure 3. Dans ce cadre, les chances sont faibles pour que les points « du futur » soient proches de cette courbe. S'en tenir seulement au degré d'ajustement du modèle aux données est donc très critiquable, que ce soit en psychologie (Roberts & Pashler, 2000) ou en EIAH (Yudelson, Pavlik Jr, & Koedinger, 2011). C'est pourquoi il est nécessaire de prendre en compte les paramètres libres du modèle.

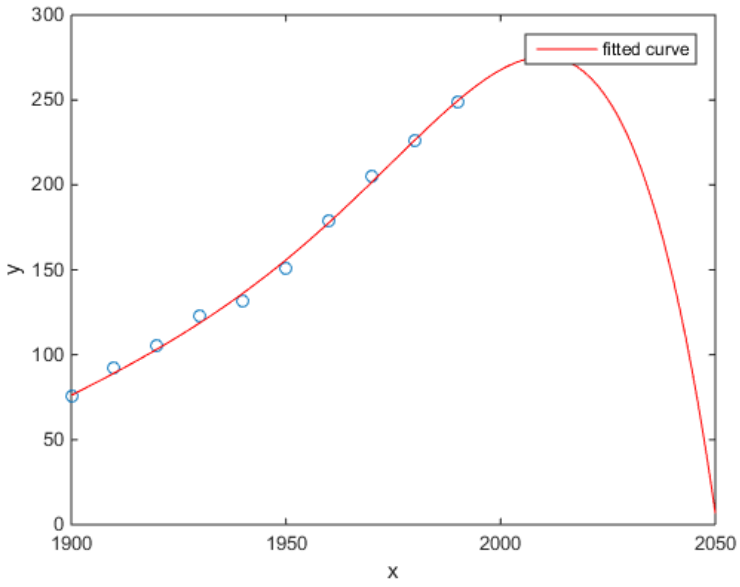



Figure 3. Interpolation de la population des États-Unis par un polynôme d'ordre 6. Malgré la proximité de la courbe aux différents points, celle-ci devient aberrante dans sa

prédiction du futur.

Source: http://www.mathworks.com/help/curvefit/examples/polynomial-curve-fitting.html#zmw57dd0e115

### 3.2.2 Importance de la notion de degré de liberté dans le cadre de la production de diagnostics cognitif.

La pertinence d'un diagnostic ne se résume pas à ses propriétés statistiques, la justification théorique du modèle de diagnostic a du poids dans la valeur des profils construits. Toutefois, s'ils ne sont pas pertinents d'un point de vue statistique, c'est signe qu'il faut revoir le modèle ou la constitution des traces sur lesquelles il s'appuie. Dans le contexte des environnements d'apprentissage, la balance entre le nombre de degrés de liberté et l'ajustement du modèle a beaucoup d'importance. En effet, les données provenant de chaque élève sont généralement limitées, ce qui augmente les risques de surajustement.

### 3.2.3 Méthodes classiques

Certains critères existent pour concilier précision et degrés de liberté d'un modèle, sous le nom de « model selection ». Les plus utilisés sont l'Akaike Information Criterion (AIC), le Bayesian Information Criterion (BIC), le Minimum Description Length (MDL) et la Cross Validation (CV). À notre connaissance, ils sont tous construits pour s'appliquer à des modèles probabilistes. Or, comme le notent certains chercheurs (Lee, 2005; Simon et al., 1995), les modèles du comportement sont souvent de nature non probabiliste. L'AIC et le BIC sont immédiatement calculables (et souvent calculés) dans les modèles employés dans les approches psychométriques, mais ce calcul n'est pas directement accessible dans les modèles non probabilistes.

## 3.3 Description du MDL

### 3.3.1 Le principe général

*« Understanding is compression! »*

Cette citation, empruntée à Chaitin (2005, p. 6), illustre de manière synthétique le principe du Minimum Description Length. Elle renvoie à l'idée générale que si l'information est comprise, alors elle peut être encodée de manière compressée. En traduisant le concept du rasoir d'Occam dans des termes mathématiques et informatiques, l'approche du minimum de complexité est transversale. C'est probablement la raison pour laquelle des chercheurs l'adoptent à la fois comme un outil

statistique, et comme un principe plus général. L'idée de compression est en effet plus étendue que son implémentation en tant que critère statistique. Elle est pour beaucoup de chercheurs une approche générale de cognition, c'est le cas de Gregory Chaitin. Ce chercheur en informatique a travaillé sur la théorie algorithmique de l'information et défend actuellement une métaphysique basée sur celle-ci. Ce courant de pensée, partagée par d'autres philosophes, porte le nom de « *digital philosophy* » (philosophie digitale). Ainsi, pour Chaitin, la compréhension du monde est équivalente à une compression de données.

Plus proche de nos considérations, l'idée que notre système cognitif peut se comprendre comme un système de compression d'information est défendue par Chater et Vitany percevant la recherche du minimum de complexité comme un principe unificateur de la cognition (Chater, 1999; Chater & Vitányi, 2003). Chater défend l'idée que sans minimum de complexité, notre capacité à extraire des motifs (patterns) du monde qui nous entoure serait mystérieuse.

Dans la suite du document, nous nous intéressons plus particulièrement à l'usage statistique de ce principe pour mesurer la qualité des modèles

### 3.3.2 Les différentes formes

Le Minimum Description Length est un principe et non une méthode unique. L'idée est de choisir le modèle qui compresse le mieux les données, mais encore faut-il définir en quoi cela consiste.

Le MDL contient quatre grandes familles : la première est le MDL idéalisé, qui correspond à la diminution de la complexité de Kolmogorov, c'est-à-dire que la taille du plus petit programme permettant de générer les données en entrée est considérée. En l'état actuel de la recherche, cette méthode n'est applicable que dans des cas particuliers (Gauvrit, Zenil, & Delahaye, 2011; Griffiths & Tenenbaum, 2003) et se base sur des approximations de cette complexité.

Les trois autres familles sont construites sur l'idée de la diminution de la complexité stochastique, ce qui est une simplification pratique de la complexité de Kolmogorov réalisée en utilisant un point de vue probabiliste. La deuxième famille est constituée des approches basées sur un codage en deux parties des données (Rissanen, 1986). La première constitue le codage des paramètres du modèle et la seconde le codage des données en utilisant le modèle. La troisième famille entend résoudre certains problèmes

liés au codage en deux parties et propose une version en seulement une partie, c'est la version « moderne » du MDL. Elle consiste à coder dans un même bloc les degrés de liberté et les données. Nous n'irons pas dans les détails de ces différentes implémentations et nous renvoyons le lecteur au livre de Grünwald (2007) pour un état de l'art général sur le sujet. Enfin, la quatrième famille existe sous le nom de Minimum Message Length (MML) développant une approche plus bayésienne du MDL. Les trois premiers types de MDL ont été utilisés dans l'étude de la cognition humaine. Le MDL idéalisé a été utilisé par certains chercheurs qui défendent l'idée que nos jugements probabilistes concernant le caractère surprenant ou non de certains évènements, peuvent être expliqués par une différence de complexité dans un sens proche de celui de Kolmogorov (Dessalles, 2013; Gauvrit et al., 2011; Griffiths & Tenenbaum, 2003). Le MDL en deux parties a été utilisé par des chercheurs comme Robinet, Lemaire, & Gordon (2011) qui proposent un modèle basé sur ce critère pour rendre compte de l'activité de chunking. Enfin, Lee et Cummins (2005) ont utilisé la version moderne du MDL pour comparer deux modèles de la décision. Une liste plus approfondie de l'approche du minimum de complexité en mathématiques/informatiques mise en lien avec ses applications en sciences cognitives peut être lue dans Chater & Vitányi (2003).

Le MDL dit « moderne » entretient une certaine proximité avec les sciences cognitives, plus particulièrement pour résoudre le cas des modèles nommés « qualitatifs » ou « déterministes ». Grunwald (1999) a mis au point une technique nommée « entropification » qui permet d'utiliser la version moderne du MDL en transformant le modèle non probabiliste en modèle probabiliste. Cette technique permet, par ce biais, d'utiliser les mêmes critères de sélection que pour les modèles probabilistes. Son principe se base sur une fonction de coût. Par exemple, s'il s'agit de prédire le temps qu'il fera demain, une fonction de coût possible est de compter +1 si la prédiction est bonne, -1 sinon. Une fois qu'elle est construite, les probabilités sont établies avec un paramètre libre qui peut être optimisé.

Si, comme le note Grunwald (1999), la distribution obtenue n'a pas de raison d'être la distribution réelle des données, elle représente la meilleure supposition accessible. Par cette technique, des modèles de nature déterministe, par exemple de décision, ont pu être comparés (Lee & Cummins, 2004).

La version moderne du MDL a des propriétés statistiques désirables, c'est pourquoi nous tenons à justifier les raisons pour lesquelles nous choisissons de travailler avec la version la plus simple du MDL qui est le code en deux parties. Les critères vus

précédemment, notamment le MDL, ne s'appliquent à des modèles déterministes que par le biais d'une fonction de coût, celle-ci permettant de construire une version probabiliste du modèle en assignant une probabilité fictive à chaque évènement. Cependant, l'évaluation d'un modèle déterministe sous une version probabiliste rend caduc le choix initial de construire un modèle déterministe. De plus, la construction d'une fonction de coût, étape cruciale, peut amener à une part d'arbitraire importante. Enfin, l'utilisation pratique du MDL « moderne » n'est pas garantie dans la mesure où des calculs assez lourds peuvent être requis pour permettre cette approche [21] (P. Grünwald, 1999). Nous notons par ailleurs que les travaux portant sur l'étude de la cognition et la sélection de modèles mettent souvent au premier plan le MDL, justement pour traiter de sélection de modèles « qualitatifs » ou « déterministes » (Lee, 2005; I. J. Myung, Pitt, & Kim, 2005; J. I. Myung, Cavagnaro, & Pitt, 2014), mais traitent ces modèles de manière probabiliste par l'utilisation d'une fonction de coût.

[I] Les calculs peuvent être simplifiés dans des cas simples (Lee, 2005)

# 4 SOURCE(S) DES ERREURS DANS LA RESOLUTION DE PAEV

Dans ce chapitre, nous analysons une partie de la littérature sur les PAEV. Dans le but de construire un modèle éclairé par la littérature sur le domaine, nous étudions plusieurs hypothèses sur la source des erreurs dans les PAEV.

## 4.1 Approche des schémas

### 4.1.1 Travaux princeps

L'avènement du paradigme cognitiviste dans les années 80 a permis de réactiver une approche dont les premières bases ont été développées en parallèle d'une période dominée par le behaviorisme : la théorie des schémas (Bartlett & Burt, 1933). Cette théorie avance que notre cognition n'est pas faite d'associations directes entre la perception et l'action comme dans le paradigme behavioriste, mais est organisée à un niveau plus abstrait et plus structuré. Le cognitivisme, cherchant à formaliser les processus et représentations mentales a donc, dans plusieurs domaines de recherche, employé et spécifié des schémas pour décrire nos activités mentales. L'approche des schémas était exploitée pour expliquer la compréhension de la lecture par les travaux du Kintsch (1983), et a ensuite été adaptée pour expliquer les processus cognitifs en jeu dans la résolution de problème arithmétique en classe élémentaire par la construction d'un programme résolvant des PAEV (Kintsch & Greeno, 1985). D'autres chercheurs,

comme Vergnaud (1976; 1982), se revendiquent plutôt de Piaget et de son concept de schème.

Prenons par exemple le problème suivant « Jean a cinq billes, mais pendant la récréation il en gagne 12. Combien de billes a-t-il maintenant ? ». Dans l'approche schéma, cette situation décrit un gain. Le schéma est donc constitué d'une quantité initiale, d'un gain (ou d'une perte), et d'une quantité finale. Ici, la quantité inconnue est la quantité finale, c'est la version la plus facile des problèmes qui mettent en jeu ce type de schéma. Les schémas ont grandement influencé ce domaine de recherche par la construction d'une typologie de problème qui capture une part importante de la différence de difficultés entre les PAEV. Les schémas sont aussi utilisés comme un modèle de compétences hiérarchiques associée à la hiérarchie de difficulté entre les différents problèmes de la typologie.

### 4.1.2 Limites théoriques et pratiques

Ces travaux servent donc de point de référence pour les approches montrant justement que les schémas « n'expliquent pas tout », et que deux problèmes appartenant au même schéma peuvent provoquer des écarts de performances importantes selon les mots utilisés (Hudson, 1983; Vicente, Orrantia, & Verschaffel, 2007 pour une synthèse) ou des variations dans les stratégies de résolution selon le choix des variables du problème (Chaillet, 2014; Gamo, Taabane, & Sander, 2011; Hakem, Sander, Labat, et al., 2005; Sander, 2007). Ces approches sont variées et ont en commun l'idée que la sémantique du problème ne se limite pas au schéma (sa structure abstraite), mais peut être décrite à un niveau plus fin, niveau dans lequel plusieurs facteurs peuvent intervenir rendant plus ou moins saillante la nature des relations mathématiques des variables du problème. Outre la sémantique du problème, la sémantique des opérations est abordée et mise en lien avec cette première. Soustraire, par exemple, n'est pas l'application aveugle de procédures abstraites, mais s'enracine dans l'acte même de soustraction comme le retrait ou le calcul de distance qui peut s'effectuer avec les doigts (ex : *« enlever 3 à 15 », « compter de 3 jusqu'à 15 »*). Comme nous l'avons dit précédemment dans le passage consacré aux aspects cognitifs de Sierra, ces deux analogies semblent expliquer les erreurs dans les soustractions posées sans même la présence d'un énoncé pour amorcer ces représentations (Sander, 2001). Dans le cadre des PAEV, si les nombres sont proches ou éloignés, une méthode de calcul (distance ou retrait respectivement) devrait être privilégiée, or c'est la manière dont le problème est représenté qui

détermine majoritairement le type de calcul qui sera effectué, ce qui peut impacter fortement la difficulté du problème (Brissiaud & Sander, 2010).

Une limite pratique de ces travaux, dans le cadre de notre thèse, est l'explication des erreurs. L'approche schéma ainsi que les nouvelles approches sémantiques porte un éclairage sur la nature des connaissances et des processus mentaux engagés dans la résolution de PAEV et fournissent des explications sur les variations de difficultés entre les différents problèmes. Toutefois, elles ne permettent pas de simuler simplement des comportements produisant des erreurs. Il est possible de construire des modèles cognitifs simulant un comportement de résolution (Briars & Larkin, 1984; Riley & Greeno, 1988), mais ces programmes se limitent à la prédiction de la réussite ou de l'échec aux problèmes. Dans la suite, nous étudierons d'autres approches, proposant des visions alternatives et complémentaires sur le traitement des problèmes par les enfants qui peuvent être employées dans la modélisation des erreurs.

## 4.2 Avec quelle profondeur les élèves traitent-ils les PAEV ?

### 4.2.1 Les questionnements soulevés par les problèmes non routiniers, réalistes ou absurdes

La résolution de problèmes arithmétiques pourrait être idéalisée ainsi : l'élève comprend le problème, se forme une représentation mentale de la situation et dérive une solution sur la base des connaissances acquises en arithmétique. Des études mettent à mal cet idéal. En effet, un certain nombre de réponses élaborées par les élèves entre en profonde contradiction avec le sens du problème. Les problèmes du type *« âge du capitaine »* recouvrent des exercices insolubles, pour lesquels la réponse n'est pas dans le reste du texte. Par exemple, à la question

*« Un berger a 12 chèvres et 6 moutons, quel âge a-t-il ? »*

Les élèves ont tendance à répondre 18. Ces problèmes ont dans un premier temps été étudiés par l'IREM de Grenoble en (1980) puis repris et réanalysés par un certain nombre de chercheurs comme Baruk (1985) et Brissiaud (1988). S'il est noté que les élèves font part de leur doute sur le problème, la réponse consistant à additionner les nombres pour obtenir la réponse reste massive. Outre ces problèmes « absurdes », les problèmes dits « réalistes » ou non routiniers (« non-routine problems ») ont également été étudiés. Un exemple classique est le problème du bus (Verschaffel, Corte, & Lasure, 1994) :

*« 450 soldiers must be bussed to the their training site. Each army bus can hold 36 soldiers. How many buses are needed? ».*

Dans ce problème, la réponse réaliste consiste à arrondir le nombre de bus à l'unité supérieure, pourtant un certain nombre d'élèves vont commettre des erreurs en ne prenant pas en compte cette conception réaliste. Les taux de bonnes réponses de ces problèmes sont très faibles. Ce phénomène dit de « suspension du sens » est d'une robustesse remarquable. Verschaffel (2010) constitue une revue importante sur cet effet et liste les pays dans lesquels les travaux de Greer (1993) et Verschaffel et al. (1994) sont répliqués : Belgique, Chine, Allemagne, Hongrie, Japon, Irlande du Nord, Suisse, Venezuela. La propension des manuels de classe et des professeurs à recourir à des problèmes stéréotypés en respectant des règles de minimalisme dans les textes de problèmes (Depaepe, De Corte, & Verschaffel, 2015; Sarrazy, 2002) provoquerait chez l'apprenant la construction et le renforcement de stratégies de résolution ne passant pas par la modélisation active de la situation problème, mais se basant plutôt sur les mots-clefs. Cette situation complexe a été formalisée de manière convergente sous différents termes. Le concept de « contrat didactique » a été proposé pour discuter des attentes réciproques des élèves et professeurs dans une situation d'apprentissage, le concept de « rationalité sociocognitive » a aussi été proposé pour discuter de la logique derrière les réponses d'élèves qui semblaient absurdes au premier abord (Reusser & Stebler, 1997). Sarrazy (2002) semble montrer que les professeurs créant des problèmes diversifiés obtiennent de meilleurs résultats dans leurs classes. Toutefois, l'établissement de contexte favorisant les réponses réalistes reste très difficile à mettre en œuvre (Reuter, Schnotz, & Rasch, 2014; Verschaffel, Dooren, Greer, & Mukhopadhyay, 2010), voire, définitivement problématique puisque la représentation fidèle du réel est impossible et que toute activité de classe ne peut ni se passer de code ni d'une médiation du langage pour désigner une situation. Il existe donc une lecture postmoderne des problèmes à énoncés verbaux selon laquelle il n'est pas possible de mettre en place de représentation transparente de la réalité (Gerofsky, 2010). Il est alors proposé de considérer

l'ambigüité des PAEV comme un obstacle indépassable et ainsi d'accepter la présence d'interprétations divergentes (Gerofsky, 1999, 2010).

L'absurdité relative des problèmes n'échappe pas non plus à la culture populaire et est souvent moquée. Une requête sur un moteur de recherche d'images comme *« the guy from your math problem »*[22] donne dans ses premiers résultats, des images à caractère humoristique exemplifiant ce que nous décrivons de manière plus académique. Nous y trouvons généralement des photographies de collections aberrantes d'un ou de deux types de consommables (bananes, ananas, ou pamplemousses...) collectés par un seul homme, dans un magasin. En Figure 4, une photographie populaire que nous reproduisons ci-contre (avec l'accord de son auteur) postée sur un site humoristique.

Figure 4 Image populaire légendée *« I am the person from your 3rd grade math word problems »*. Cette photographie à caractère humoristique porte sur l'artificialité des problèmes en classe élémentaire.

### 4.2.2 Critique de l'hypothèse de lecture superficielle

Les études précédentes suggèrent parfois que les élèves sont dans une attitude superficielle face aux problèmes qui leur sont posés, parfois appelée stratégie basée sur

[22] Sauvegarde de la recherche : http://archive.is/xip3g

les mots-clefs (Hegarty, Mayer, & Green, 1992). Par cette stratégie, les élèves ne modélisent pas mentalement la situation décrite dans le problème et se basent sur des mots-clefs pour deviner l'opération nécessaire. Cette hypothèse a été renforcée par les études portant sur la résolution des problèmes réalistes, avec l'idée que (1) l'échec des élèves résulte de l'absence de modélisation mentale et que (2) ces stratégies permettent de résoudre les problèmes stéréotypés auxquels ils ont l'habitude d'être confrontés. Cependant, cette hypothèse des mots-clefs exclut les facteurs sémantiques comme composante explicative des erreurs. Or, la robustesse de ces effets documentés dans un corpus de taille toujours plus grande ne peut que rentrer en contradiction avec cette idée. Par ailleurs, l'idée selon laquelle seuls les élèves en difficultés recourent à ce genre de stratégie est en contradiction avec les données de la plupart des études sur les problèmes réalistes. Il est en effet montré que même si les élèves en difficulté résolvent moins bien les problèmes pièges, le taux d'échec chez les « bons » élèves reste plus qu'important[23]. De ce fait, considérer la stratégie de résolution par mot-clef comme seule explication des échecs aux problèmes réalistes ne peut pas suffire pour rendre compte du taux d'échec élevé. Il s'agit donc de nuancer la portée de cette hypothèse en l'intégrant aux autres composantes connues de la résolution de PAEV. Par ailleurs, Inoue (2005) montre qu'une activité interprétative peut avoir lieu sur ces problèmes dits « réalistes » : une partie des élèves interrogés semble avoir réussi à construire une représentation du problème qui leur permet d'aboutir à la réponse — bien qu'erronée — la plus stéréotypée. Cette étude peut être rejointe par celle de Brissiaud (1988) qui note que la grande majorité des élèves détectent bien que les problèmes représentent des anomalies (20/23). II est difficile d'intégrer ces nouvelles connaissances en modélisation et de les faire interagir avec d'autres composantes de la résolution de problème, car elles portent sur des niveaux d'abstraction différents. Nous verrons que la flexibilité du formalisme du modèle des contraintes permet de concilier ces approches.

[23] De tels résultats peuvent être retrouvés, par exemple, pour l'exercice 18 dans (Nortvedt, 2011)

### 4.2.3 Recontextualisation des comportements dits superficiels par un système de contraintes

S'il est tentant de qualifier ces comportements de résolution comme irrationnels du point de vue de la situation représentée de l'énoncé, il est possible de défendre au contraire une rationalité de l'apprenant dans son adaptation au problème réel, qui dépasse le cadre de la situation décrite, mais qui réside dans la résolution de situation de classe et dans lesquels ses connaissances mathématiques sont mêlées aux connaissances concernant le « genre particulier » que constituent les PAEV (Gerofsky, 1999). Donner une réponse qui *a priori* a de bonnes chances d'être juste est rationnel dans ce cadre. Ainsi, De Corte, Verschaffel, & De Win (1985) considèrent qu'un ensemble de connaissances générales sur les PAEV doit être acquis pour pouvoir les résoudre, dénommé « Word Problem Schema ». Ainsi, sans ces connaissances *a priori* sur la forme des problèmes posés, le problème suivant, qui ne contient qu'un nombre, n'est trouvé surprenant que par 6 élèves parmi 31 (selon De Corte et Verschaffel,1983, cité dans Brissiaud 1988) « Pierre avait des pommes ; il a donné 4 pommes à Anne ; combien de pommes a-t-il maintenant ? » Toujours chez ces auteurs, l'aspect artificiel des PAEV est parfois décrit par le terme « Word Problem Game » pour représenter les compétences spécifiques et décalées de l'application au monde réel, qui sont incidemment construites. Gerofsky (1996) liste un ensemble étendu de présuppositions sur les PAEV :

- Le problème est résoluble.
- L'ensemble des informations permettant la résolution est dans l'énoncé.
- La résolution doit être effectuée par le biais d'un algorithme mathématique précédemment enseigné.
- Il n'y a qu'une seule interprétation du problème et qu'une seule réponse juste, que le professeur détient.

Ces présuppositions peuvent aussi être vues comme des connaissances nécessaires à la résolution de la plupart des problèmes tels qu'ils sont présentés dans un cadre scolaire. Une autre manière de percevoir cet effet est de le voir sous l'angle de la communication humaine qui respecte un certain nombre de contraintes. Notamment, il est supputé que l'interlocuteur donne au locuteur l'exacte quantité d'information nécessaire. Paul Grice, père de la linguistique pragmatique, a introduit le terme d'« implicatures conversationnelles » pour qualifier les inférences qui en sont tirées. De ce fait si un

interlocuteur indique que *« Jacques a rencontré Pierre ou Paul. »* Il est naturel dans le cadre de la conversation humaine d'en déduire que Jacques n'a pas rencontré Pierre ET Paul, car si tel était le cas, notre interlocuteur l'aurait dit. De même, comme l'indiquent Kintsch & Greeno (1985), reprenant les propos de Pearla Nesher & Katriel (1977), la phrase *« Jean a 8 billes »* signifie dans le cadre d'un PAEV *« Jean a <u>exactement</u> huit billes »* et non pas *« Jean a au moins huit billes »* qui dans un autre contexte pourrait être une interprétation tout à fait correcte. Ils montrent aussi que le sens des propositions est absolument restreint aux seules quantités, lorsque *« Jean donne 5 billes à Paul »* on ne se questionne pas sur les raisons de ce don, ni sur ses conséquences comme la gratitude de Paul. Le fil narratif dans lequel s'insère cette proposition n'est pas celui d'une histoire « classique ».

Il est donc visible qu'un énoncé est compris en prenant en compte l'intention du locuteur, la cohabitation des deux implique que le sens donné à un énoncé ne coïncide pas avec une interprétation formelle du texte du problème. Cette règle de qualité implique qu'un interlocuteur est censé maximiser l'information transmise, donc éviter du contenu superflu. Dans le cadre des PAEV, cette règle de communication justifierait la tendance des élèves à vouloir utiliser tous les nombres de l'énoncé. Cette règle, appelée « règle de complexi*té maximale »* par Danièle Coquin-Viennot (2001), peut être mise en évidence en demandant aux élèves de compléter un texte de problème dans lequel il manque la question. Coquin-Viennot (2000) montre que les propositions les plus fréquentes sont celles qui mettent en jeu toutes les quantités du problème. Coquin-Viennot (2001) observe aussi une tendance à employer toutes les données de l'énoncé alors que ce n'est pas toujours nécessaire, ce qui laisse penser qu'ils ne répondent pas à la question posée, mais à la question qui semble la plus typique pour l'énoncé. Cette question inférée peut par ailleurs être plus difficile que la question réelle du problème. Il est supposé que les élèves, par habitude, portent une attention moindre à la question qu'aux autres éléments du texte, du simple fait qu'elle porte généralement sur une quantité déductible de toutes les informations du texte.

Ceci pose un problème dans la modélisation des processus de résolution. Ces phénomènes ont été très étudiés, mais isolément d'autres aspects de la résolution et ne font pas partie intégrante des modèles proposant d'expliquer les erreurs.

### 4.2.4 Support pour la contrainte « ne pas réutiliser de nombre »

À cette règle, nous pouvons en ajouter une seconde qui lui est proche: « ne pas utiliser deux fois un nombre de l'énoncé ». Cette contrainte peut être mise en lien avec la représentation des PAEV suggérée par Palm (2006) qui considère que les problèmes sont vus comme un simple habillage d'une tache mathématique. On peut suggérer que la règle « 2 nombres impliquent une opération » trouve support dans cette conception. S'il n'existe pas de données de la littérature, certaines expérimentations fournissent des résultats allant fortement dans ce sens. Plus précisément, lorsque cette contrainte doit être relâchée pour obtenir de bonnes réponses, les performances sont généralement mauvaises. C'est ce que nous pouvons retrouver dans les données de Castro-Martínez & Frías-Zorilla (2013). En effet ils ont employé des problèmes dont la bonne réponse viole cette règle d'utilisation des nombres. Ils faisaient l'hypothèse et ont montré un effet de la structure du problème sur sa difficulté. Deux types de problèmes sont proposés selon le nombre d'éléments communs entre les deux schémas qui les constituent (cf. Figure 5).

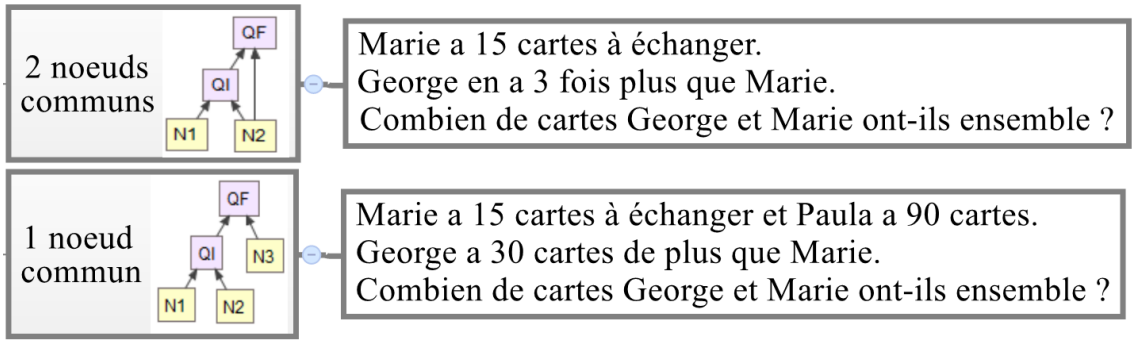



Figure 5. Exemples de problèmes à un ou deux nœuds communs. Problèmes issus de Castro-Martínez & Frías-Zorilla (2013) traduits en français. L'illustration est la nôtre.

Les problèmes ayant un nœud « commun » dans leurs structures sont plus difficiles (Tableau 2), les erreurs sont très largement des erreurs étiquetées *« une seule opération »*. Le nombre important d'erreurs de ce type dans les problèmes de la première catégorie, même s'il est relevé par les auteurs, reste inexpliqué. En effet, l'étude conclut (p. 404) sur :

> *« This study could be continued by tackling from a qualitative point of view the psychological reasons for the different student errors in problems with one and two nodes. »*

En étudiant la partie « *matériel* » de l'article, il apparait que ces changements de structure impliquent aussi une modification notable portant sur les nombres de l'énoncé. En effet, pour qu'un problème soit une instance de la structure « avec nœud commun », seulement deux nombres doivent être présentés. Toutefois, le mode de résolution du problème ne varie pas, deux calculs sont nécessaires. Ceci implique que la réutilisation d'un nombre pour obtenir la réponse est nécessaire pour résoudre les problèmes de ce type. Lorsqu'une seule opération est effectuée, les erreurs sont divisées en deux parties : oubli de la première ou de la deuxième relation du problème. Leurs données détaillées représentent pourtant un atout puissant pour argumenter en faveur de cette contrainte, car les chercheurs ont établi une catégorisation sur les types d'erreurs qui contiennent la catégorie « *une seule opération* ». Si la contrainte « *ne pas réutiliser de nombre* » a un poids fort, alors cette erreur devrait être fréquente dans les problèmes « à nœud commun ». Pour quantifier cet effet, nous avons repris les données de cette étude en les résumant sur un tableau 2x2 sous la forme « erreur, car un seul calcul » et « autres erreurs », nous obtenons le Tableau 2.

Tableau 2 Réorganisation des données de Castro-Martínez & Frías-Zorilla (2013). Mise en correspondance du type de structure de problème et du type d'erreur.

| Type de problème | Type d'erreurs | |
|---|---|---|
| | Une seule opération | Autres erreurs |
| 1 nœud commun | 119 | 9 |
| 2 nœuds communs | 4 | 56 |

L'erreur « une seule opération » représente 93 % des erreurs obtenues pour les problèmes avec deux nombres donnés (1 nœud commun). Au contraire l'erreur « une seule opération » ne représente que 7 % pour l'autre catégorie (2 nœuds communs, 3 nombres donnés). Tester notre hypothèse nulle portant sur l'inexistence de la contrainte de réutilisation des nombres revient à faire un test d'indépendance entre les deux problèmes. La statistique suivante est obtenue : $X^2 = 130.7$, dl = 1, p-value < 1.10-4. Cette statistique présente l'inconvénient de ne pas quantifier l'effet. Nous choisissons donc de tester le rapport de proportion par la méthode de Katz, Baptista, Azen, & Pike (1978). L'intervalle de confiance à 95 % du ratio est compris entre 3.8 et 12.8. Qu'en est-il des adultes ? Réinterprètent-ils les problèmes pour se conformer à une résolution stéréotypée ? C'est une manière d'interpréter les réponses à ce problème :

> *Un bonbon et une baguette coûtent 1,10 euros ensemble. La baguette coûte un euro de plus que le bonbon. Combien coûte le bonbon ?*

La plupart des adultes interrogés répondent « *10 centimes* », ce qui n'est pas la bonne réponse (5ct). À notre connaissance, ce problème, actuellement très célèbre en psychologie du raisonnement (Bourgeois-Gironde & Van Der Henst, 2009; De Neys, Rossi, & Houdé, 2013), ne fait pas beaucoup d'écho dans la psychologie des apprentissages arithmétiques au niveau élémentaire. L'idée sous-jacente est qu'il s'agit simplement d'un problème piège qui n'apporte pas d'information sur l'apprentissage de l'arithmétique, car l'adulte « a déjà appris ». Le champ conceptuel associé à ce problème est celui des biais cognitifs et de la rationalité humaine, en témoigne cette citation (De Neys et al., 2013, p. 3) :

> *After all, the problem is really not that hard. Clearly, if people would reflect upon it for even a moment they would surely realize their error and notice that a 10 cents ball and a bat that costs a dollar more cannot total to $1.10*

Or, une interprétation de la difficulté du problème guidée par nos analyses précédentes peut replacer ce problème dans le domaine de la résolution des PAEV. Tout d'abord, en ce qui concerne l'utilisation des nombres de l'énoncé, il n'est pas pensable que le problème requière plus d'une opération, encore moins une équation ((1+x)+x=1.10). C'est donc de manière très intuitive que la réponse 10ct survient. Ensuite, on peut aussi supposer que la phrase seule *« La baguette coûte un euro de plus que le bonbon »* ne pose pas de problème particulier de compréhension. Une suggestion possible est que la proximité de la phrase *« la baguette coûte un euro de plus que le bonbon »* avec la phrase *« la baguette coûte un euro »* joue un rôle dans le processus. Dans le langage des contraintes, il pourrait être dit que la contrainte d'intégrité sémantique de la première phrase a été relâchée. Pour reprendre les termes des défenseurs de l'hypothèse de résolution superficielle, si les adultes se construisent un modèle du problème, dans ce cas ils ne devraient pas échouer.

## 4.3 Expliquer les difficultés par le traitement des éléments du problème par l'oculométrie.

### 4.3.1 Forme des études employant l'oculométrie

La résolution de PAEV a été étudiée sous l'angle de l'oculométrie. Le but de ces études est d'apporter un nouveau point de vue sur les stratégies de résolutions et la cause de

certaines erreurs en testant des hypothèses sur le traitement des éléments du problème. Cette méthode permet par exemple de séparer le temps de première lecture et le temps de relecture, ainsi que de s'intéresser aux parties spécifiques du texte sur lesquelles se pose le regard des bons élèves et les élèves moins performants (De Corte, Verschaffel, & Pauwels, 1990; Hegarty et al., 1992, 1995; van der Schoot, Bakker Arkema, Horsley, & van Lieshout, 2009). Ces études sont importantes dans le cadre de nos recherches sur la résolution pas-à-pas de problème arithmétiques. Plus précisément, il serait discutable de suivre la théorie selon laquelle l'élève favorise des stratégies sémantiques plutôt que superficielles, si les données en oculométrie suggèrent une prédominance de ces dernières. Peu d'études en eye-tracking ont été réalisées dans ce domaine, nous en sélectionnons cinq, portant justement sur la question de la superficialité des stratégies de résolution de l'élève. Ces cinq études sont portées par trois équipes différentes : van der Schoot et al (2009), Verschaffel et De Corte (1990, 1992), Hegarty et Mayer (1992, 1995). Nous nous référons aussi au travail de Verschaffel (1994), car même si ce n'est pas une étude en oculométrie à proprement parler, sa problématique est en lien avec les études susnommées et enregistre les temps de lecture. Les données issues de ces travaux sont relativement convergentes. Outre l'oculométrie, toutes ces études ont le point commun d'avoir un matériel décomposé en (1) problèmes concordants et problèmes discordants et (2) formulation « de plus » ou « de moins ».

Comme présenté en Tableau 3, l'opération permettant la résolution de ces problèmes est sémantiquement opposée à la relation décrite dans l'énoncé : il faut faire une soustraction dans les problèmes qui contiennent l'expression « de plus ».

Tableau 3. Exemple de problèmes concordants et discordants traduits de (Verschaffel, De Corte, & Pauwels, 1992) .

| Type d'opération demandée | Problème concordant | Problème discordant |
|---|---|---|
| Addition | Brian a 32 livres.<br>Ralph a 13 livres de plus que Brian.<br>Combien Ralph a-t-il de livres ? | Marie a 38 bagues.<br>Elle a 25 bagues de moins que Joan.<br>Combien de bagues a Joan ? |
| Soustraction | Carol a 35 poupées.<br>Ann a 29 poupées de moins que Carol.<br>Combien de poupées a Ann ? | Pete a 28 crayons .<br>Il a 17 crayons de de plus que Dick<br>Combien de crayons a Dick ? |

Plusieurs interprétations de ce phénomène peuvent être relevées dans la littérature, dont l'hypothèse dite de Lewis & Mayer (1987). Il est proposé que l'erreur provienne d'un échec d'une opération mentale de renversement de la relation de comparaison. Même si

trois publications sur les six effectuent ces études sous l'hypothèse de Lewis et Mayer, comme noté par Verschaffel, le matériel ne permet pas facilement de la différencier de l'hypothèse de lecture superficielle, qui se trouve donc elle est aussi discutée. Cette stratégie est basée sur les mots-clefs et les nombres du problème pour essayer de deviner les opérations demandées, elle ne cherche pas à construire un modèle mental de la situation décrite dans le problème.

### 4.3.2 Hypothèse de traitement superficiel des problèmes discutée

Outre l'argumentation des auteurs, de fortes convergences dans les données présentées permettent de discuter l'idée d'une lecture superficielle. En effet, l'idée d'un traitement superficiel du problème est souvent défendue (Hegarty et al., 1992; Hegarty et al.,1995), avancée, mais nuancée (Verschaffel 1992) ou fortement nuancée, voire critiquée (Van Der Schoot 2009, De Corte 1990, Verschaffel 1994). Une analyse attentive des résultats présentés dans ces études nous mène à un point de vue nuancé, car malgré le degré variable de soutien des différents chercheurs à l'hypothèse d'une lecture superficielle, les résultats présentent de bonnes convergences. Si les bons et mauvais élèves semblent dépenser autant de temps dans la lecture, la relecture est plus longue chez les élèves en difficulté, et ce sur l'ensemble des problèmes.

### 4.3.3 Analyse des temps de lecture

En ce qui concerne l'effet de la discordance des problèmes, les élèves vont généralement dépenser plus de temps dans la lecture et la relecture d'un problème discordant que dans un problème concordant, et ce, autant pour les bons élèves que pour les élèves en difficulté. Ils sont donc sensibles à la complexité des relations dans le problème. L'étude de Verschaffel (1994), qui enregistre les temps de résolutions sans oculométrie, confirme aussi cet aspect. Les temps de lecture (ou de relecture De Corte et al. 1990) sont généralement plus longs pour les élèves en difficultés. Par ailleurs, lorsque la difficulté du problème augmente, les temps de lecture sont plus longs, ce qui va dans le sens d'un traitement sémantique des problèmes donnés. Il y a cependant une exception notable, dans l'étude de van Der Schoot et al. (2009), des durées de résolutions anormalement courtes sont observées pour certains problèmes discordants accompagnés des performances très mauvaises pour le groupe en difficulté. Ces problèmes sont les problèmes dits « non marqués », car ils contiennent la formule « moins de ». Ce pattern pourrait effectivement être la trace de lectures superficielles. Ainsi, les conclusions sont donc nuancées (p. 58) :

> *« The conclusion was drawn that, like more successful problem solvers, less successful problem solvers can appeal to a problem-model strategy, but that they do so only when the relational term is unmarked. »*

Les données de Hegarty et al. (1992) vont aussi dans ce sens : même si la version « moins de » est lue autant sinon moins vite que la version « plus de » chez les élèves en difficulté, on note toutefois que la version discordante « moins de » est lue beaucoup plus lentement par les bons élèves. Autrement dit, la règle empirique selon laquelle plus un problème est difficile, plus les temps de relecture sont longs fait exception ici, ce qui peut en effet suggérer un traitement superficiel des problèmes. Ainsi, même si l'effet de la discordance sur les temps de lecture s'aligne avec les autres études et montre que l'enfant n'est pas directement dans une lecture superficielle, ce pattern isolé montre qu'avec une difficulté supplémentaire (« moins de » à la place de « plus de ») la construction d'un modèle de la situation échoue « plus rapidement » chez les mauvais élèves. Cette idée de pression supplémentaire peut être mise en correspondance avec Verschaffel (1992) qui montre qu'il est difficile de mettre en évidence l'effet du marquage « plus de » à la place de « moins de » chez les étudiants à l'université sur des problèmes simples, mais qu'il est visible pour des problèmes complexes.

### 4.3.4 Types d'éléments observés

Dans l'étape de relecture, le traitement des éléments semble différer entre les bons et les mauvais élèves. Les mauvais élèves semblent — en proportion — plus concentrer leur attention sur les nombres et les mots indiquant des relations que sur les variables du problème (De Corte et al., 1990; Hegarty et al., 1995). Cette différence est un argument pour l'existence de lectures superficielles. Enfin, une étude de notre synthèse propose un argument supplémentaire pour les stratégies superficielles. Pour rechercher si les élèves en difficultés ont une approche plus superficielle que les bons élèves, Hegarty et al. (1995) ont complété leurs études avec une tâche de rappel des énoncés soumis aux élèves. Les problèmes discordants étaient rappelés avec un taux important d'erreurs sémantiques, surtout chez les moins bons élèves. Autrement dit, les élèves en difficulté commettent plus d'erreurs altérant le sens du problème, ce qui serait en faveur de stratégies superficielles selon les auteurs. Cependant, l'altération du sens du problème pourrait être aussi bien la marque d'une réinterprétation lors de la résolution ou d'une erreur d'inversion selon l'hypothèse de Lewis Mayer. En effet, l'étude de Verschaffel (1994), similaire à l'étude de Hegarty dans la mesure où des problèmes concordants et

discordants sont rappelés, montre que les erreurs de rappel sont en fort accord avec la représentation du problème sous l'hypothèse d'inversion de Lewis et Mayer, donc en lien avec la représentation formée par le sujet. Devant ces résultats, les auteurs concluent (p. 159) :

> *« For these two reasons, we stick to our first assumption, namely, that the large majority of the pupils in the present study did not apply the superficial key-word strategy when solving compare problems. ».*

Voir les erreurs de rappel comme la marque d'une mauvaise compréhension basée sur une fragilité linguistique est une interprétation possible, comme nous allons le voir au travers des travaux de Cummins. S'il est possible de noter au travers de ces études quelques éléments en faveur d'un traitement superficiel des problèmes, les données de ces études sont plutôt en faveur d'un traitement profond des problèmes. Ainsi, la plupart des études concluent sur l'idée que la stratégie par mots-clefs n'est pas suffisante pour expliquer le comportement des élèves en difficulté, et qu'un traitement sémantique — ou une tentative de traitement — a bien lieu chez ces élèves. Bien que Hegarty et al. (1992, 1995) ont tendance à soutenir l'hypothèse superficielle, les résultats obtenus diffèrent peu des autres études analysées.

## 4.4 Les difficultés dans la lecture : Un important facteur explicatif des erreurs

### 4.4.1 L'importance des capacités de lecture et de compréhension

La résolution superficielle peut être vue non pas comme une cause, mais comme un symptôme des difficultés de résolution rencontrées par l'élève. Nous avons déjà abordé la théorie des schémas et le modèle des compétences qui en résulte, il est également possible d'analyser les aspects linguistiques de la résolution de problème. Pour qu'un enfant active et emploie un schéma permettant la résolution, il doit comprendre la situation à laquelle il est confronté, ce qui nécessite des connaissances linguistiques. Si l'énoncé n'active pas ces schémas, alors le processus de solution ne peut pas aboutir selon les chercheurs du camp linguistique. Dans ce cadre, les schémas, considérés par certains chercheurs comme nécessaires à la résolution sont acquis plus précocement qu'il n'y paraît, mais la maitrise incomplète du langage chez les jeunes enfants ne permet pas d'en tirer parti. Selon cette dernière hypothèse, les schémas pourraient être acquis relativement tôt sans pour cela pouvoir être employés efficacement, car la

situation décrite dans le problème n'est pas bien comprise. Les premiers travaux en « rewording » (reformulation) de Hudson (1983) montrent qu'une version reformulée d'un même problème obtient des performances radicalement différentes alors que le même schéma est mis en jeu. Des chercheurs comme Cummins (1988) voient dans ces études un support pour l'importance des compétences linguistiques.

Plusieurs preuves peuvent être réunies soulignant l'importance de la capacité de lecture dans la résolution de PAEV. Tout d'abord, des mesures de corrélation entre les scores obtenus dans des épreuves mathématiques et les épreuves langagières sont souvent établies. Si des mesures globales de corrélation sont un bon indice pour explorer une question de recherche, elles restent évidemment à être expliquées par des études plus précises. Elles peuvent aussi être contrôlées par d'autres mesures. Swanson, Cooney, & Brock (1993) montrent une relation entre la mémoire de travail et les performances en PAEV, mais cette relation est significative seulement si la capacité de compréhension de texte, meilleur prédicteur, est reléguée de force au second plan. Les travaux menés par Fuchs étudient les corrélations entre diverses compétences et la capacité de résolution de PAEV. Multiplier ainsi les facteurs permet de hiérarchiser les corrélations entre les différentes capacités et de rendre compte d'éventuels effets de médiation. Fuchs et al. (2006, 2012) reportent que la capacité de lecture explique une variance « unique » au sein d'autres facteurs comme la capacité de calcul, l'attention ou la mémoire de travail. Il en ressort que les capacités de lecture sont un bon corrélat des capacités des résolutions de problèmes (plus que la capacité à effectuer des opérations par exemple).

### 4.4.2 Une capacité spécifique de compréhension des PAEV mise en avant

Fuchs et collaborateurs (Fuchs, Fuchs, Compton, Hamlett, & Wang, 2015) montrent que les scores de compréhension de texte corrèlent avec les performances dans les PAEV et que la compréhension du **langage spécifique** employé dans les problèmes représente une médiation moyenne de cet effet. Cette analyse peut être mise en correspondance avec celles proposées précédemment poussant à voir les PAEV comme un genre littéraire particulier. Schumacher & Fuchs (2012) montrent l'efficacité d'une intervention basée sur la maitrise du vocabulaire spécifique aux PAEV. En outre, une analyse de médiation a confirmé l'idée que l'amélioration obtenue varie avec l'amélioration de la compréhension de ces termes. Par ailleurs de nombreuses études proposent et mesurent l'impact des alternatives à la présentation de l'énoncé sous

formes textuelles seules. Elles soulignent le fait que l'activité de lecture peut être un blocage de taille à la résolution de problème.

L'idée que les PAEV posent des problèmes de compréhension qui leur sont spécifiques est renforcée par l'étude de Vilenius-Tuohimaa (2008), qui en dissociant capacité de lecture technique et compréhension de la lecture, montre que chacun de ces aspects explique une partie des performances. Roe & Taube (2004) sont en accord avec ces résultats dans leur étude sur les performances au test PISA. La capacité de compréhension est trouvée plus explicative des performances dans les PAEV en plusieurs étapes que la capacité technique de lecture. Une étude longitudinale a aussi permis de montrer que la capacité de compréhension est un bon corrélat des compétences en résolution de PAEV.

### 4.4.3 La méthodologie et l'apport de Cummins sur les aspects linguistiques

Parmi les travaux les plus poussés sur les aspects linguistiques de la résolution de problème arithmétique, nous devons mentionner et étudier les travaux de Denise Dellarosa Cummins (Cummins, 1991; Cummins et al., 1988). Elle fait partie du camp « linguistique », opposé aux approches développementales défendues par Riley & Greeno (1988) et Briars & Larkin (1984). Elle utilise (1988, 1991) des tâches de rappel, de création de questions et de sélections d'images représentant le problème pour montrer que les erreurs sont en accord avec les difficultés de compréhension des PAEV. Les performances aux tâches de rappel sont cohérentes avec les performances précédentes de ces mêmes problèmes, si on les explique par des interprétations erronées.

Une autre dimension importante des recherches de Cummins (1988) est l'utilisation d'un programme informatique pour simuler des réponses. Notre thèse prenant le parti pris de la modélisation, nous étudions en détail son approche. Trois modèles sont proposés, ils sont tous la représentation d'une lacune dans les capacités de résolution de problème. La problématique est de déterminer quel est le plus prédictif des réponses observées. Ces modèles sont obtenus après dégradation du modèle de résolution de problème par schéma de Briars et Larkins (1984). Le premier ne dispose pas de schémas de haut niveau permettant de conceptualiser les relations entre les quantités du problème. Par conséquent, il opère une traduction directe des mots du problème en opération. Il correspond donc au comportement hypothétique des élèves les moins performants décrits dans notre passage sur l'hypothèse superficielle. Le deuxième modèle comprend les mots du problème, mais n'a pas les compétences de

compréhension de l'histoire. Il se laisse guider par les mots-clefs non pas pour inférer des opérations comme pourrait le faire le premier modèle, mais pour inférer des schémas. Le troisième modèle, soutenant la position théorique de Cummins, est défaillant du point de vue linguistique sans atteinte des deux autres compétences (schémas et compréhension). Il interprète de manière incorrecte des mots-clefs tels que « quelques » (Some), « plus que » (More than) et « ensemble » (Altogether). Le caractère exécutable des modèles permet de dériver des prédictions sur chacun des problèmes de la typologie de Riley et Greeno (1988). Elles sont uniques (une prédiction par problème) et peuvent être erronées : mauvaise opération, pas d'opération ou bonne réponse (bonne opération). Cette prédiction est alors comparée à la réponse la plus fréquente des élèves pour chacun des problèmes. Un score allant de 0 (aucune prédiction vérifiée) à 18 (toutes les prédictions sont vérifiées) est donc calculé. Le troisième modèle (lacunes linguistiques) réalise 15 prédictions correctes sur 18 le premier 5 sur 18 et le deuxième 7 sur 18. Ces résultats sont donc en faveur du modèle de défaillance linguistique.

Nous retenons la démarche qui consiste à dériver d'un modèle des prédictions en termes de réponses attendues. Elle permet de tester des hypothèses de manière plus ciblée qu'une expérimentation qui ferait l'hypothèse d'une variation de performance selon la formulation du problème. Nous notons aussi que le modèle de résolution superficielle semble être le moins explicatif des trois, ce qui conforte notre position décrite précédemment.

La technique consistant à établir une prédiction unique pour ensuite la confronter à la réponse la plus fréquente des élèves présente un intérêt pratique notable puisqu'il existe peu de réponses possibles à un PAEV à une seule étape (à quelques exceptions près : additionner, soustraire ou ne rien faire). Mais il faut noter, d'une part, que ces réponses majoritaires sont établies après récolte auprès d'une faible quantité d'élèves, d'autre part, que l'analyse des réponses non majoritaire n'est pas réalisée.

# 5 MODELISATION DU COMPORTEMENT DANS LES EIAH PORTANT SUR L'ARITHMETIQUE

## 5.1 Introduction

L'objectif de ce chapitre est de fournir une revue sur les environnements et les outils visant l'apprentissage de la résolution de problèmes arithmétiques à énoncés verbaux (PAEV) en classe élémentaire et d'interroger leur relation avec la psychologie cognitive. Ce chapitre reprend assez largement l'article de colloque que nous avons publié dans la conférence EIAH 2015 (Bruno Martin, Labat, & Sander, 2015). Nous conservons la trame générale de l'article qui pose des grandes catégories situant ces outils les uns par rapport à autres que nous complèterons par une analyse des aspects cognitifs de ces outils dans le contexte particulier de la recherche en psychologie dans le domaine des PAEV. Nous nous centrons sur les outils qui portent sur la conceptualisation et la résolution de PAEV. Les environnements doivent montrer un effort sur ce domaine et documenter leurs réalisations pour pouvoir être inclus dans une synthèse. Pour ces raisons, nous éliminons de notre analyse les « suites commerciales ». Ce sont des programmes qui proposent une série d'activités mathématiques couvrant une large gamme de niveaux scolaires (parfois de la classe primaire à la fin du lycée). Ainsi les PAEV ne représentent qu'une faible partie de ces environnements, non documentée et probablement peu élaborée. Pour des raisons différentes, nous écartons

les programmes qui visent à améliorer l'appréhension du nombre et l'estimation des quantités chez les enfants. Basés sur les travaux sur le sens des nombres, les programmes tels que Calcularis (Käser et al., 2013), ou la course aux nombres (Wilson et al., 2006), visent à entraîner l'enfant à des estimations de quantité, en visant particulièrement les enfants dits dyscalculiques. S'il est possible que cette aptitude joue un rôle dans les apprentissages mathématiques comme la résolution de PAEV, la force du lien est sujette à débat. Par ailleurs ces programmes ne portent pas sur la résolution de PAEV, c'est pourquoi nous ne les détaillerons pas dans l'analyse.

## 5.2 Critères de catégorisation des outils

Nous avons identifié deux stratégies majeures et deux stratégies mineures employées pour construire cet apprentissage. La première est la décomposition, plus ou moins poussée, de la résolution en différentes étapes. La deuxième est l'utilisation de représentations externes, plus ou moins abstraites, du problème. Ces deux stratégies vont généralement de pair : plus la décomposition est marquée, plus les représentations proposées à l'apprenant sont concrètes. La Figure 6, qui rassemble les outils de cette double catégorisation, représente cette relation de manière qualitative. Les environnements sont représentés dans un plan constitué de deux axes : la nature des représentations engagées dans la tâche et le degré d'autonomie laissé au sujet.

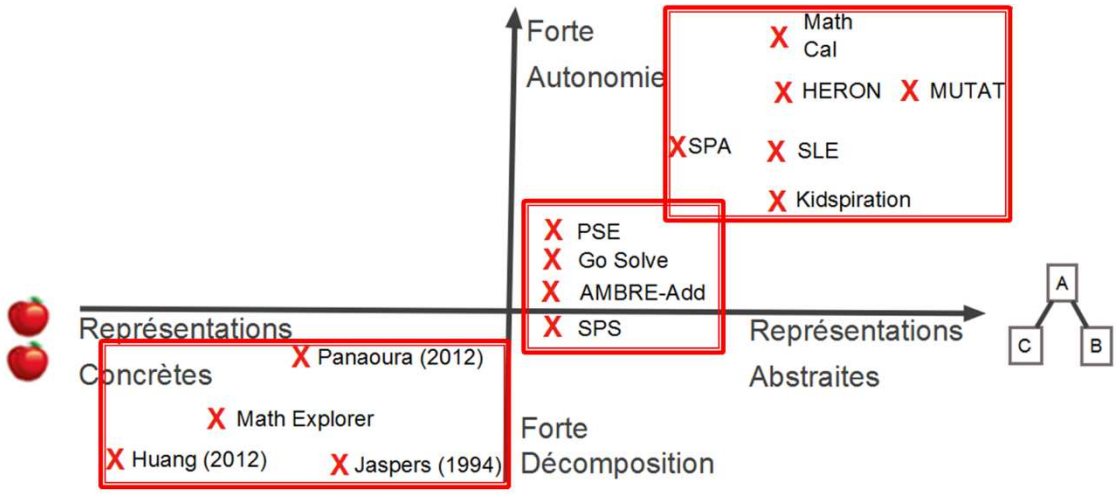



Figure 6. Présentation d'une partie des environnements d'apprentissage dans le domaine de l'arithmétique en classe élémentaire.

D'autres stratégies, moins fréquentes, sont aussi employées dans quelques environnements : (i) chercher à modéliser ou évaluer l'apprenant, et (ii) proposer des activités détournées de la résolution directe de problèmes. Ces stratégies seront dites

« mineures » pour leur plus petite représentation et seront traitées dans des parties à part.

### 5.2.1 Représentations graphiques intermédiaires

La grande majorité des environnements d'apprentissage portant sur les PAEV invite le sujet à visualiser, sélectionner, compléter ou même produire une représentation graphique du problème. En effet, puisque la résolution de PAEV ne semble pas pouvoir faire l'économie de la notion de représentation, l'idée de proposer à l'apprenant une représentation graphique du problème paraît s'imposer.

Ces représentations graphiques peuvent être classées selon leur niveau d'abstraction par rapport à la situation problème. Ainsi le dessin représentant les objets (et éventuellement les acteurs du problème) est le niveau le plus concret alors que l'écriture de la relation mathématique est le plus abstrait. Entre les deux se situent les représentations de collections dans lesquelles les objets représentés sont abstraits (des carrés par exemple) et les modélisations schématiques qui caractérisent des quantités sans représenter de collection. Encore plus abstrait, car éloigné de la notion de représentation de la situation, se situent les arbres de solutions et les schémas présents dans HERON (Reusser, 1993).

### 5.2.2 Décomposition en étapes

Deux approches sont souvent citées pour justifier la décomposition observée dans les différents outils. La première, s'appliquant à la résolution de problèmes en général, est celle de Polya (1945). Les étapes proposées sont (1) comprendre le problème (2) établir un plan (3) suivre ce plan (4) vérifier les résultats. Plus spécifique à la résolution de PAEV, Verschaffel (De Corte et al., 2000; Verschaffel, 2000) est invoqué, de manière moins fréquente cependant. Les étapes proposées par Verschaffel sont (1) comprendre la situation, (2) la modéliser mathématiquement, (3) dériver des résultats, (4) évaluer ces résultats en regard de la situation et (5) conclure. L'approche de Polya est assez générale, celle de Verschaffel est plus spécifique aux problèmes arithmétiques en mettant au centre la conceptualisation et la modélisation de la situation. L'établissement d'un plan ne fait pas partie de la décomposition de Verschaffel, car ce dernier défend l'idée que la résolution de PAEV doit reposer sur une activité de modélisation mentale du problème pour permettre sa résolution. Les outils en questions utilisent parfois la méthode du chercheur en faisant travailler les enfants avec des problèmes non

stéréotypés qui comportent des données manquantes ou au contraire inutiles pour les inviter à adopter une démarche active de modélisation (Jaspers & Van Lieshout, 1994; Panaoura, 2012).

## 5.3 Catégorisation des outils basés sur les stratégies majeures

### 5.3.1 Faible décomposition et représentations abstraites : outils pour les problèmes difficiles

Au rang des outils les plus influents se situe HERON, conçu par Reusser (1993). Il se réclame de Nathan (Nathan, Kintsch, & Young, 1990) en suivant l'idée de construire un tuteur non intelligent (unintelligent tutor). Ces tuteurs se définissent par la parcimonie sur la modélisation du problème ou de l'élève et se concentrent sur l'outillage de l'apprenant pour l'aider dans la résolution. En effet, il ne s'agit pas pour Reusser de chercher à modéliser finement l'élève, mais de lui fournir un environnement de résolution riche qui lui serve d'échafaudage pour construire sa solution. Le cœur de sa proposition est l'utilisation d'arbres de solution (solution tree, cf. Figure 7), constitués de schémas triadiques pour lesquels les quantités nouvellement calculées se trouvent en bas. Chaque nœud du schéma correspond à une quantité associée à une valeur, un label désignant ce qu'elle représente et une unité (litres par exemple). C'est à l'utilisateur de sélectionner des nombres et d'y associer ces informations. Ces quantités peuvent être glissées dans des schémas, eux-mêmes glissés dans la fenêtre de résolution. Cette manière de représenter les solutions est particulièrement adaptée aux problèmes qui nécessitent plusieurs étapes de calculs. De plus, la généricité de l'outil et des schémas employés le rendent utilisable pour un ensemble large de problèmes (dont les multiplications et divisions). HERON est un environnement de résolution qualifié par son auteur de constructiviste modéré, car l'enfant est libre de la construction de son arbre et de la suite d'actions par lesquelles il passe. Il peut par exemple passer d'une fenêtre à une autre pour récupérer de nouvelles quantités dans le texte ou continuer son arbre de solution. HERON ne propose pas toute l'architecture d'un tuteur intelligent, mais l'environnement fonctionnant par descriptions de quantités et par choix d'opérateur permet, si l'enfant le demande, d'obtenir des feedbacks sur sa progression et des notifications d'erreurs.

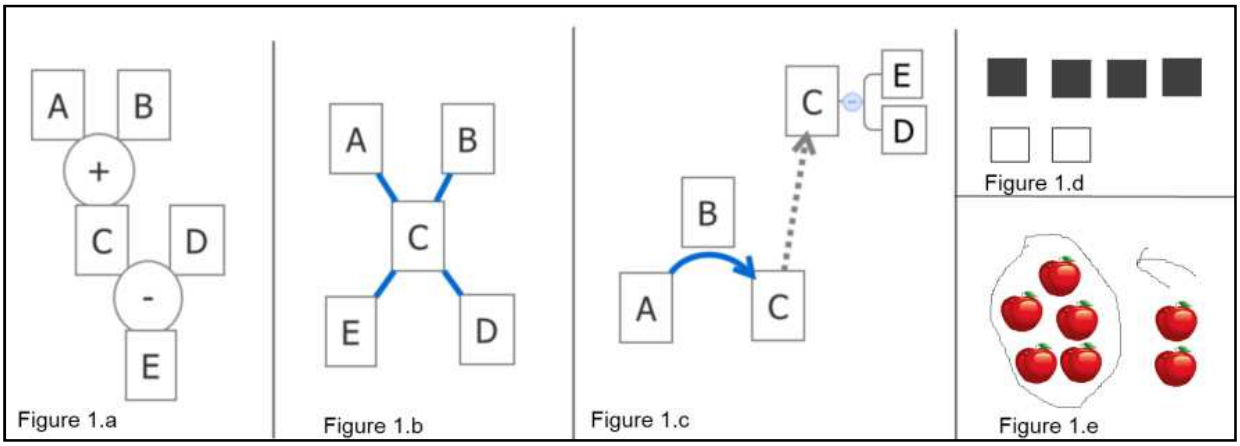



Figure 7 Différentes représentations des quantités et situations selon les différents outils.

1.a : Arbre de résolution adapté de Reusser (Reusser, 1993).
1.b : Schème adapté de SPA, le tout C est partagé (Hershkovitz & Nesher, 1998).
1.c : Schémas adaptés de SPS (Marshall, 1995)
1.d : Collections abstraites, adaptées de Jaspers et Van Lieshout (Jaspers & Van Lieshout, 1994)
1.e : Collections concrètes et dessin adaptés de Math Explorer (Seo & Woo, 2010)

Deux programmes inspirés par HERON ont été développés. Le premier, Math Cal (Chang, Sung, & Lin, 2006), est porté par des chercheurs qui regrettent que HERON ne force pas la décomposition étape par étape. Deux arguments sont avancés. Le premier concerne le diagnostic, avec l'idée qu'il est plus facile de comprendre les erreurs en segmentant les activités. Le deuxième concerne la charge mentale. Décomposer les étapes permettrait d'alléger la mémoire de travail de l'apprenant. Ainsi, dans Math Cal, une fois les quantités définies par l'utilisateur, une étape explicite de planification est réalisée, dans laquelle l'apprenant construit une liste ordonnée des quantités à résoudre.

Le deuxième programme alternatif, peut-être plus riche, est le Solver Learning Environnement (Ruokamo & Pohjolainen, 1998). Ce dernier s'inscrit dans une approche web et multimédia. Des éléments du problème sont cliquables et permettent ainsi d'activer différents types d'aides comme l'accès à des définitions. L'élève peut aussi visualiser des vidéos représentant le problème pour l'aider. Enfin, l'aspect web permet aux élèves d'adopter une démarche collaborative. S'ils sont en difficulté, ils peuvent accéder à une fenêtre leur permettant de discuter avec les autres élèves ou avec le professeur.

À défaut de voir cet aspect clairement mentionné par les auteurs de ces deux programmes, il semble que le diagnostic ne soit possible qu'à la fin du processus, en comparant la solution de l'élève avec une solution enregistrée, ce qui constitue un appauvrissement du programme original. En effet HERON indiquait la présence d'incohérences au cours de la résolution. C'est peut-être justifiable pour Solver Learning Environnement du fait de sa dimension socioconstructiviste, mais plus discutable pour Math Cal qui cherche justement à guider l'élève pas-à-pas.

Schemes for Problems Analysis (SPA) (Hershkovitz & Nesher, 1998), tout comme Math Cal, met en question le fait que HERON autorise l'apprenant à ajouter des quantités au fur et à mesure. La raison est cependant différente de Math Cal ; pour Nesher et Hershkovitz, il est plutôt question de représentation des relations entre quantités. Selon ces auteurs, la possibilité d'ajouter des quantités au fil de l'eau freine une conceptualisation de la structure globale du problème. Dans leur logique, la première action de l'élève devrait être de choisir les schémas qui décrivent la situation. L'utilisation des schémas dans HERON est alors qualifiée de rhétorique puisque le schéma ne met en relation que des quantités déjà conceptualisées par l'enfant. La représentation d'un schème est graphiquement semblable aux schémas dans l'arbre de solution de HERON. Cependant les quantités ont une place différente selon leur statut. Un schème est composé de deux parties et d'un tout. Dans SPA, les apprenants sélectionnent des schèmes (multiplicatifs ou additifs), pour ensuite former un schème composite. Trois sont possibles selon que (1) le tout est partagé par les deux schèmes, c'est le schème représenté en Fig. 1.b, C étant le tout partagé ; (2) la partie d'un schème est partagée par les deux schèmes ; et (3) la partie d'un schème est le tout d'un autre. Une fois le schème composite construit, les enfants peuvent résoudre le problème en indiquant les nombres et labels dans chaque schème pour ensuite calculer la quantité inconnue.

Une évolution conceptuelle est notable de HERON à SPA. En effet, dans SPA, la relation entre les quantités est prioritaire à leurs définitions. La représentation graphique change ainsi de rôle, elle n'est plus utilisée pour calculer des quantités inconnues sur la base de quantités connues, mais elle sert à représenter ces quantités elles-mêmes.

Le principe de HERON de faire glisser des quantités dans un organisateur graphique a été repris par des outils fonctionnant de manière tactile comme dans Multi-Touch Arithmetic Tool (MuTAT) (Adesina, Stone, Batmaz, & Jones, 2014) et Kidspiration 3©(Sheriff & Boon, 2014). Ce dernier propose de glisser les quantités dans des schémas

préétablis dont les labels sont déjà présents. La place des nombres et l'opérateur sont ensuite choisis par l'enfant. MuTAT est encore plus épuré ; les nombres sont glissés sans qu'il soit question de les intégrer dans un schéma spécifique. Cependant il permet de construire des arbres de solution comme dans HERON.

### 5.3.2 Décomposition modérée et représentations semi-abstraites

Situation Problem Solver (SPS), développé par Marshall (1995), contient un ensemble de leçons qui permet à l'apprenant d'associer une situation problème à un schéma pour aborder sa résolution. C'est une forme assistée par ordinateur d'instruction par des schémas (Fuchs et al., 2004). Selon Marshall, ces derniers sont de cinq types : transformer, combiner, comparer pour les problèmes additifs, varier et réitérer pour les problèmes multiplicatifs. Chacun de ces schémas fait l'objet d'une représentation graphique particulière. Sur la base de ce programme, un environnement de résolution permettant à l'apprenant de résoudre librement des problèmes a été conçu sous le nom de Problem Solving Environnement (PSE) (Marshall, 1995). Proche de SPA, cet environnement s'en distingue puisque l'apprenant glisse des schémas spécifiques à l'intérieur de la fenêtre principale. Les quantités identiques sont reliées par des flèches (Fig 1.c). Un autre chercheur, Derry s'est également appuyé sur l'approche schéma. Son élaboration a été accompagnée d'un environnement, TAPS (Training Arithmetic Problem-solving Skills) (Derry, Hawkes, & Tsai, 1987) qui deviendra par la suite TIPS (Tutorial in Problem Solving) (Lajoie & Derry, 2013).

Par l'emploi de schémas similaires à ceux présents dans SPS, certains outils se concentrent sur la résolution de problèmes simples. Ils proposent généralement une décomposition des étapes, accompagnant l'apprenant par des aides et des feedbacks. GO Solve Word Problems (Leh & Jitendra, 2013) est un programme qui décompose la résolution en plusieurs étapes : Sélection d'un schéma, insertion des quantités, et écriture de la solution. GO Solve Word Problems peut aussi être utilisé pour les problèmes nécessitant plusieurs calculs, mais sur la base de schémas simples « étendus » avec un quatrième argument et non composites comme sur SPA.

Ambre-add (Nogry, Guin, & Jean-Daubias, 2008) est un projet similaire qui invite l'apprenant à choisir un schéma et à le compléter avec les quantités de l'énoncé. Une fois cette étape réalisée, la particularité d'Ambre est d'inviter l'apprenant à résoudre le problème par analogie avec un exercice de même structure. Cette étape dérive de l'approche dite « raisonnement à partir de cas » qui défend l'idée que les nouveaux

problèmes sont résolus par analogie avec des problèmes précédemment résolus (Aamodt & Plaza, 1994).

Huang et collaborateurs ont conçu un outil utilisant aussi les schémas, qui adopte également une décomposition forte des étapes de résolution, avec des aides graphiques accessibles à chaque étape (Huang, Liu, & Chang, 2012). La première étape consiste à choisir une reformulation de la question parmi trois proposées. S'il clique sur l'aide, l'enfant peut observer un dessin représentant les quantités des problèmes. La deuxième étape propose de choisir une équation et non plus un schéma. Le schéma n'est donc ici plus une représentation imposée ; il apparait seulement si l'étudiant demande l'aide pour cette étape. Comme dernière étape, il est laissé la possibilité à l'enfant de conduire et vérifier son calcul. Cette stratégie de tutorat qui décompose finement les étapes de résolution en s'assurant de la bonne compréhension de l'apprenant, a été développée dans d'autres outils que nous détaillons dans la partie suivante. Ils visent, tout comme l'outil de Yang et collaborateur, les élèves en difficulté.

### 5.3.3 Décomposition forte et représentations concrètes : outils pour les élèves en difficulté

En ce qui concerne les élèves ayant des difficultés, il n'est pas surprenant que le schéma soit supplanté par la manipulation de représentations plus concrètes accompagnées d'activités centrées sur la compréhension et la modélisation de la situation du problème. L'outil développé par Jaspers et Van Lieshout (1994) est un représentant de cette approche. L'interface est assez simple, mais il se concentre sur des activités clefs finement décomposées : l'attention au texte, la modélisation active de la situation et la manipulation de représentations concrètes. Les problèmes comportent des données inutiles pour conduire l'apprenant à modéliser la situation, et les nombres sont petits pour que le calcul ne soit pas un obstacle. Après avoir lu une première fois le problème et sélectionné les mots importants de la question, l'élève effectue trois activités sur chacune des phrases : la relire, sélectionner les mots et les quantités dont le rôle est important, et enfin les représenter par une collection de petits carrés (Fig 1.d) en les sélectionnant dans l'interface.

Une fois que les deux ensembles sont représentés, il est demandé à l'enfant de donner une réponse numérique à la question. Avec une interface plus moderne et aussi destinée aux élèves en difficulté, Math Explorer (Seo & Woo, 2010) décompose la résolution en quatre étapes, chacune suivie d'invitations à des réflexions métacognitives (« J*'ai bien*

*fait X* » et « *Je vérifie que mon étape est bien réalisée* »). La première étape est la lecture, la deuxième le dessin, et la troisième le calcul. L'étape de dessin présente la spécificité d'inviter l'enfant à utiliser des outils dans une mini-interface pour réaliser son croquis l'outil crayon lui permet d'entourer à la main des quantités, et l'outil image de glisser-déposer des icônes représentant des éléments du problème (Fig 1.e).

Un autre outil très visuel a été développé par Panaoura (2012). Ce programme se décompose en leçons animées par un cartoon, qui de manière similaire aux autres programmes, apprend à l'enfant une suite d'étapes. Comme pour l'outil de Jaspers et Van Lieshout, le but est d'amener l'enfant à construire un modèle de la situation. Sa première étape est assez originale, car elle consiste à se poser des questions sur la situation problème, à la résumer, à mettre de côté des données inutiles ou, au contraire, à remarquer qu'il manque des données et à les demander au programme pour résoudre le problème.

## 5.4 Alternatives aux tuteurs

Dans cette dernière partie, nous abordons les programmes qui ne passent pas par l'emploi d'un environnement de résolution riche ou par la décomposition pas-à-pas de la résolution de problème.

### 5.4.1 Modéliser et diagnostiquer

Merlin's Math Mill (Schoppek & Tulis, 2010) et AnimalWatch (Cohen, Beal, & Adams, 2008) sont deux outils dont l'interaction avec l'utilisateur est minimale (présence d'un seul champ de réponse et très peu d'interaction avec le tuteur). Les efforts des chercheurs dans ces projets se sont portés sur la modélisation de l'apprenant. Invoquant le concept de zone proximale de développement de Vygotski (1978), l'idée est de pouvoir poser le bon problème au bon moment grâce au diagnostic réalisé par ces programmes. Ces deux modèles proposent des manières originales de modéliser l'apprenant. AnimalWatch a mis en place un modèle de l'apprenant conçu à partir de régressions linéaires cherchant à prédire le temps et la qualité des réponses pour permettre à un module nommé ADVISOR d'optimiser le choix des problèmes (Beck, Woolf, & Beal, 2000). Les facteurs utilisés dans les régressions linéaires ne sont pas très détaillés, il est indiqué qu'ils sont au nombre de 48 et appartiennent à 4 catégories : (1) le développement cognitif de l'apprenant, (2) la difficulté du type d'exercice, (3) la

difficulté de l'exercice particulier et (4) le contexte (efforts de l'apprenant sur la question courante, ainsi que les indices qui sont donnés).

La modélisation mise en place dans Merlin's Math Mill utilise des méthodes plus « symboliques », elle se base sur la relation de prérequis. Pour pouvoir résoudre certains problèmes, il faut être en mesure de résoudre des problèmes du niveau inférieur. La hiérarchie des compétences proposée (nommée HiSkA pour polyhierarchy of skills in arithmetic) intègre les problèmes à énoncés verbaux, les problèmes de calcul simples (5+3= ?), ainsi que les procédures de calcul. La typologie des PAEV proposée par Riley et Greeno est utilisée et étendue aux activités de calculs et aux problèmes à plusieurs étapes. Le niveau de granularité de cette hiérarchie de compétence est donc d'un niveau de granularité plutôt fin. Nous notons que les auteurs ont fait un effort important pour comparer les résultats statistiques de cette approche avec les approches plus classiques tel le modèle de Rash qui est une version simple des régressions logistiques dans l'Item Response Theory (Schoppek & Landgraf, 2011).

TAPS, que nous avons présenté dans la partie précédente, pourrait aussi être cité ici dans la mesure où des travaux de modélisation par le biais de réseaux de neurones ont été mis en place pour modéliser les capacités de planification de l'apprenant (Posey & Hawkes, 1996).

ECBM (Web-based Curriculum-Based Measurement ; Tsuei, 2007) et Accelerated Math (Atkins, 2005) sont des outils qui permettent non plus d'intervenir sur le niveau de l'apprenant, mais d'établir un diagnostic à l'attention du professeur. ECBM et Accelerated Math sont des outils qui permettent au professeur de suivre le niveau mathématique de sa classe.

### 5.4.2 Autres outils pour aborder la résolution de problème

Les outils TIMA (Tools for Interactive Mathematical Activity ; Steffe & Olive, 2002) et la plateforme constructiviste de Garcia et Pacheco (Garcia & Pacheco, 2013) sont issus de ce courant. TIMA est un environnement purement constructiviste dans lequel les enfants manipulent des collections d'objets. Dans le deuxième outil, la manipulation d'objets répond à une question arithmétique donnée dans l'interface comme « ajouter 2 poissons ». Dans la mesure où ces plateformes ne portent pas directement sur la résolution de problèmes arithmétiques, elles ne sont pas classées dans les stratégies principales.

MONSAKUN (Hirashima, Yokoyama, Okamoto, & Takeuchi, 2007), l'outil de Chang et collaborateurs (Chang, Wu, Weng, & Sung, 2012), ainsi qu'un module d'AnimalWatch (Birch & Beal, 2008) visent à assister la création de problème par des élèves. Cette activité est reconnue comme ayant des effets bénéfiques sur l'apprentissage et la motivation des élèves pour les PAEV (Akay & Boz, 2010). En effet, lorsqu'un enfant pose un problème, il doit se construire un modèle mental et le traduire dans un texte de problème. Ces outils permettent d'assister la création d'un problème. Ils diffèrent dans la manière dont ils appuient et utilisent les problèmes construits par l'élève, mais les détailler serait hors de l'objet de cette synthèse.

## 5.5 Bilan de la synthèse du point de vue de la psychologie cognitive

### 5.5.1 Prépondérance de la notion de schémas dans les outils de cette synthèse.

Les chercheurs à l'origine de HERON et SPA se fondent sur leurs propres études théoriques et expérimentales (Pearla Nesher & Hershkovitz, 1994; Reusser, 1990), inspirés des travaux de Kintsch et Van Dijk (Kintsch & Van Dijk, 1978) utilisant le concept de schéma pour modéliser la compréhension de texte. Une généalogie décrivant les supports théoriques des environnements impulsés par des chercheurs en psychologie peut être établie, nous la représentons en Figure 8.

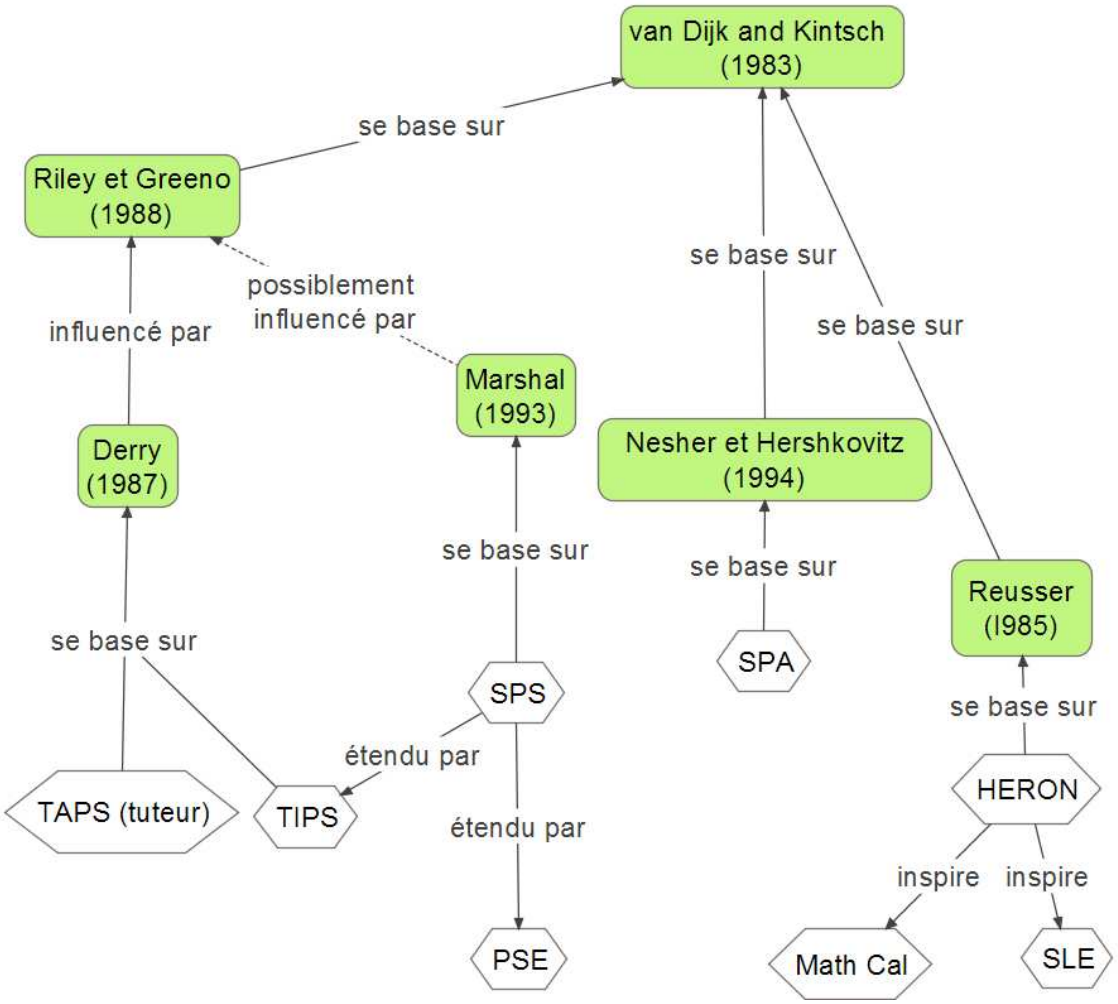



Figure 8. Généalogie des schémas dans les EIAH liés à l'arithmétique. En vert sont représentés les travaux théoriques ou expérimentaux des psychologues ayant développés les EIAH représentés par les liens directs aux EIAH sont représentés par des hexagones.

Comme nous l'avons indiqué dans la partie théorique, l'approche de Kintsch a été complétée avec la collaboration de Greeno (Kintsch & Greeno, 1985) pour expliquer les difficultés des PAEV à une seule étape. Exception faite des outils tels que HERON et SPA portés par des variations théoriques proposées par leurs auteurs, la typologie des schémas dans les EIAH suit cette approche. Nous n'avons donc pas représenté en Figure 8 tous les environnements se basant sur la typologie de Riley et Greeno, mais avons conservé seulement ceux développés par des psychologues.

L'approche « schéma » dont se réclame une partie des environnements étudiés a été en effet influente en matière d'enseignement (Schema Based Instruction ; Fuchs et al., 2004 ; Willis & Fuson, 1988). Elle consiste à demander à l'enfant de sélectionner et de travailler sur le schéma qui se rapproche le plus du problème. Ainsi, un certain nombre d'outils ont proposé des représentations graphiques pour assister la résolution de

problème. Ces schémas sont aussi utilisés dans des outils assistant la résolution de PAEV à plusieurs étapes représentant alors une alternative à HERON, dans lesquels ils sont encore plus « concrets » que SPA.

Or, comme nous l'avons souligné dans notre partie théorique, la recherche dans ce domaine a été certes marquée par l'utilisation des schémas, mais ne s'y arrête néanmoins pas. Si la théorie des schémas a profondément influencé la recherche et l'enseignement des mathématiques, la représentation d'un problème ne se résume pas à sa structure sous-jacente (Vicente et al., 2007). Des éléments plus contextuels peuvent influencer fortement sa résolution. Son niveau de difficulté peut être largement modulé par la manière dont la situation du problème est décrite. L'utilisation de ce type de connaissances pourrait améliorer des programmes comme Merlin's Math Mill ou Animal Watch qui cherchent à optimiser le choix des problèmes à proposer. Enfin nous notons que malgré le fait que beaucoup d'EIAH du domaine sont développés par des psychologues, ils visent un apprentissage construit sur la base des conceptions théoriques de ses auteurs, mais ne construisent pas des EIAH construits pour permettre à des théories d'évoluer.

### 5.5.2 Limites de la modélisation dans les environnements étudiés.

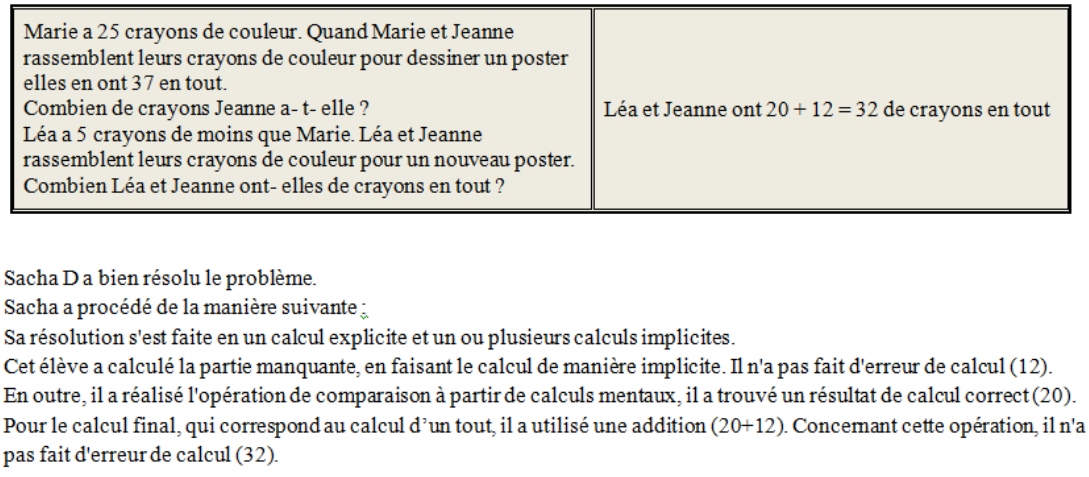
S'il est vrai que certains outils comme Animal Watch ou surtout Merlin's Math Mill élaborent des modèles assez évolués des compétences de l'apprenant, les modélisations dépassent rarement l'évaluation binaire des productions (réussite/échec). Trois difficultés surviennent cependant lors de la réalisation de ces analyses : (1) leurs détections/catégorisations, car détecter une erreur spécifique est plus compliqué à détecter que le résultat n'est pas celui attendu, et la difficulté est d'autant plus importante que les données sont constituées de réponses ouvertes ; (2) leurs interprétations, car elles impliquent un parti pris théorique en psychologie qui doit être justifié ; enfin (3) les remédiations, car, conformément au *« don't diagnose what you can't treat »* de Self, avoir détecté et interprété une erreur ne garantit pas de savoir comment la traiter. Dans le cadre de notre thèse, les deux premiers points représentent une partie de nos travaux, le troisième point, comme nous l'avons suggéré précédemment, est laissé de côté.

# 5.6 Le projet DIANE

## 5.6.1 Description du projet

### 5.6.1.1 Fonctionnalités et objectif de DIANE

DIANE est l'acronyme de Diagnostic Informatique de l'Arithmétique au Niveau Élémentaire. Son principe est d'établir des diagnostics fins des réponses ouvertes à des PAEV. Cette spécificité permet à DIANE de se distinguer des outils décrits dans l'état de l'art qui précède. Nous l'avons justement retiré dans la synthèse qui précède pour lui dédier une partie plus détaillée. Actuellement, l'interaction avec l'élève est minimale, dans la mesure où son activité se limite à la résolution de problèmes. La seule rétroaction dépendant de la réponse de l'élève est un message d'erreur si celle-ci est laissée vide, il est alors invité à donner une réponse. La recherche d'interaction (aides, tutorat, scénarisation adaptative…etc.) n'est pas la priorité de l'environnement. DIANE est constituée de deux interfaces : une interface élève et une interface administrateur (Hakem, Chaillet, & Sander, 2011; Hakem, Sander, Labat, et al., 2005). Nous les détaillons ci-après. Le diagnostic est construit comme un système de description fine des protocoles. Nous parlons alors de « codage » de la réponse de l'élève. Dans DIANE, ces indicateurs sont représentés par des colonnes de codage et leurs valeurs sont représentées par des chiffres. Nous décrivons et analysons ce diagnostic plus loin dans le document. Le codage brut est difficile à lire, même pour les experts, car il est nécessaire de se souvenir de la signification de chaque chiffre pour les différentes colonnes. De ce fait, un retour en langage naturel (cf. Figure 9) a été conçu pour traduire ce diagnostic d'une manière compréhensible par le chercheur et le professeur.



| | |
|---|---|
| Marie a 25 crayons de couleur. Quand Marie et Jeanne rassemblent leurs crayons de couleur pour dessiner un poster elles en ont 37 en tout.<br>Combien de crayons Jeanne a- t- elle ?<br>Léa a 5 crayons de moins que Marie. Léa et Jeanne rassemblent leurs crayons de couleur pour un nouveau poster.<br>Combien Léa et Jeanne ont- elles de crayons en tout ? | Léa et Jeanne ont 20 + 12 = 32 de crayons en tout |

Sacha D a bien résolu le problème.
Sacha a procédé de la manière suivante :
Sa résolution s'est faite en un calcul explicite et un ou plusieurs calculs implicites.
Cet élève a calculé la partie manquante, en faisant le calcul de manière implicite. Il n'a pas fait d'erreur de calcul (12).
En outre, il a réalisé l'opération de comparaison à partir de calculs mentaux, il a trouvé un résultat de calcul correct (20).
Pour le calcul final, qui correspond au calcul d'un tout, il a utilisé une addition (20+12). Concernant cette opération, il n'a pas fait d'erreur de calcul (32).

Figure 9. Description du diagnostic en langage naturel.

L'interface « élève » permet à ce dernier de disposer d'un espace de travail à partir duquel il peut répondre à des problèmes arithmétiques. La Figure 10 est une capture d'écran de cette interface. Le cadre supérieur gauche est celui de l'énoncé, inférieur gauche le brouillon, et à droite la copie. Le programme s'adressant à de jeunes enfants (niveau classe élémentaire), cette interface se veut facile d'utilisation. L'élève peut utiliser le pavé numérique virtuel pour écrire ses calculs et cliquer sur les mots de l'énoncé pour formuler sa réponse. Ainsi, même sans maitrise de son clavier d'ordinateur, l'élève peut répondre seulement à l'aide de sa souris. Une fois le brouillon rempli, l'élève peut alors cliquer sur « écrire dans la feuille » et si son résultat lui convient, cliquer sur « exercice terminé ». S'affiche ensuite l'exercice suivant.

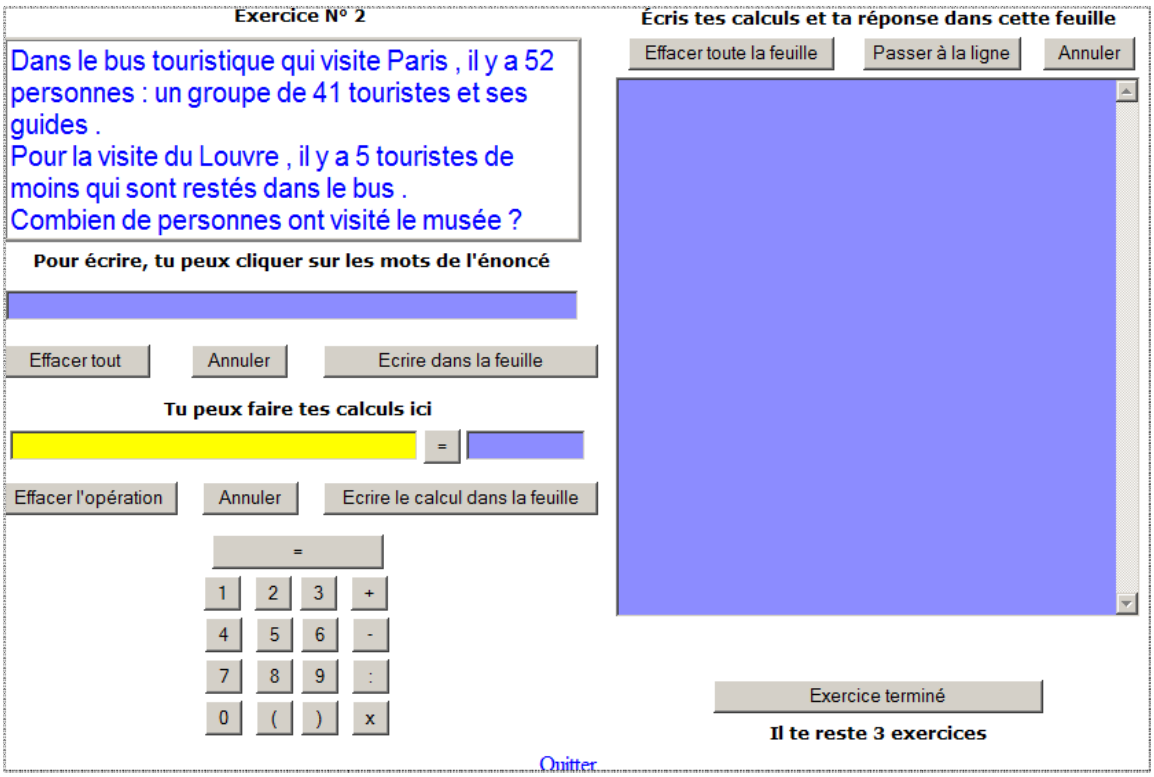



Figure 10. Interface de résolution.

5.6.1.2 Création de problèmes dans DIANE

Lorsqu'un professeur ou un chercheur veut rajouter un problème dans DIANE, il doit d'abord sélectionner sa catégorie d'appartenance. Actuellement, les types proposés sont : problème à une étape, problème complexe, problème de distributivité. Chaque type de problème a son interface création. En Figure 11, est représentée l'interface pour la création d'un problème additif à une étape. L'utilisateur doit ensuite renseigner les différents champs. Par exemple, s'il choisit « *problème à une étape* », il détermine ces propriétés fondamentales selon la typologie de Riley et Greeno en sélectionnant :

- Son schéma sémantique (combinaison, comparaison, transformation)
- La place de l'inconnue dans la relation mathématique (ex. : Etat initial, Etat final, Gain).

Un exemple est ensuite proposé dans un encart pour éclairer l'utilisateur sur la forme du problème attendu lorsque ces deux caractéristiques sont choisies.

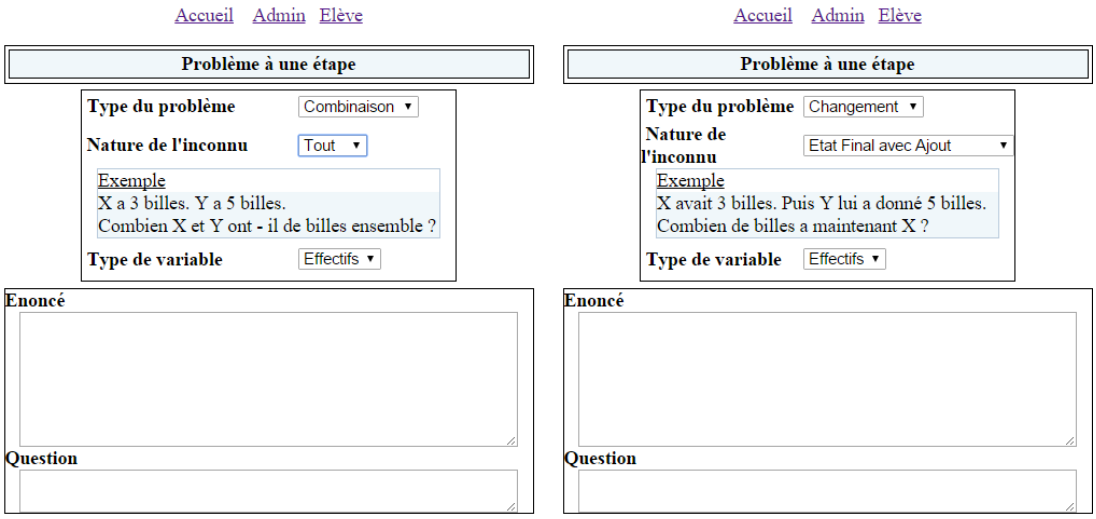



Figure 11. Deux exemples de l'utilisation de l'interface de création de problèmes simples.

La création de problèmes « complexes » est décrite en Figure 12. Les problèmes dits complexes sont des problèmes nécessitent plusieurs étapes pour être résolus. Certains champs sont optionnels ou ont peu d'importance dans la forme actuelle du diagnostic de DIANE (ordre des données, charge mentale, type de contexte, tendance, stratégie de résolution), mais peuvent jouer un rôle dans les analyses statistiques réalisées postérieurement à la récolte des données.

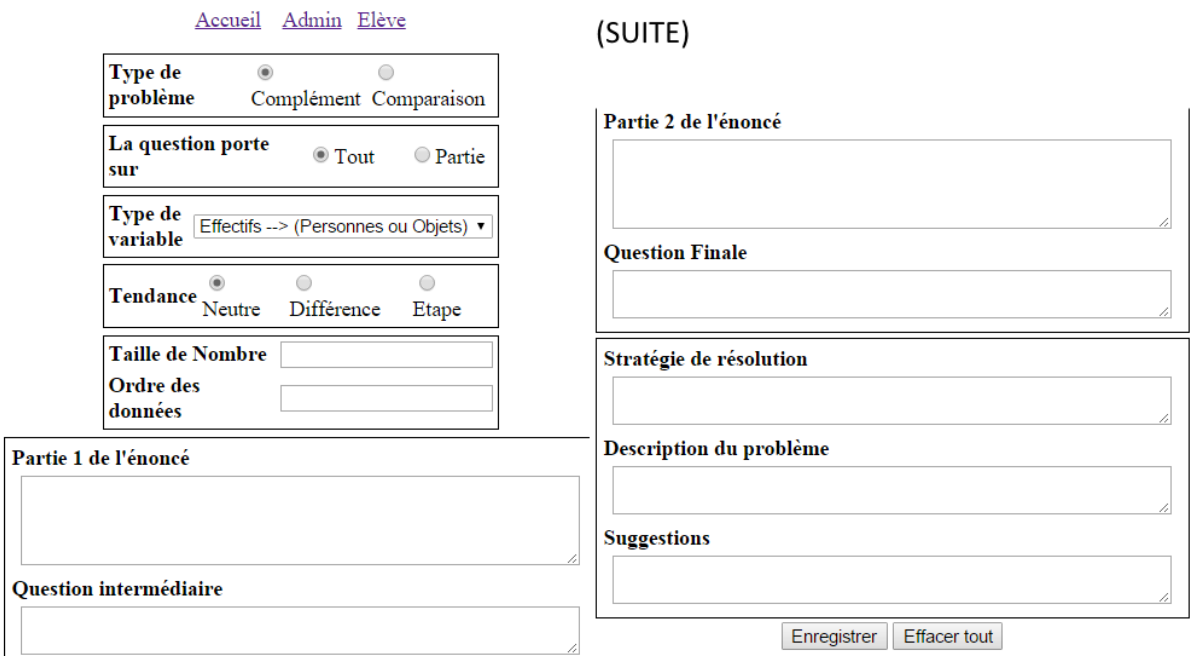



Figure 12. Interface de création de problèmes complexes.

Une fois le problème créé, il peut être sélectionné par un professeur dans le cadre d'une série d'exercices accessible lorsque l'élève se connecte à l'application.

5.6.1.3 Diagnostic des problèmes complexes dans DIANE

En tant qu'outil de recherche pour les psychologues, DIANE rejoint l'idée de mariage interdisciplinaire que nous proposions en introduction. Dans la synthèse qui précède, beaucoup d'outils sont développés par des psychologues, ne sont pas considérés comme un moyen pour mener des expérimentations et tester des hypothèses en psychologie cognitive. DIANE a été utilisée pour analyser les facteurs pouvant influer les stratégies de résolutions de problèmes complexes. Un certain nombre de recherches proposent que les quantités peuvent avoir une représentation ordinale (durées, hauteurs) ou une représentation cardinale (nombre d'éléments d'une collection) et que cette représentation influence les stratégies de résolution (Gamo, Nogry, & Sander, 2014; Gamo, Sander, & Richard, 2010). DIANE a été utilisée dans le recueil de données et l'analyse automatique de protocoles par une grille d'indicateurs permettant de pister si les stratégies de résolutions attendues ont été observées de manière totale (découverte de la bonne réponse) ou partielle (abandon ou erreurs). Les réponses à un certain type de problèmes, appelés « problèmes complexes » ont donc été analysée. Ces problèmes ont la particularité intéressante d'être tous solubles en une ou trois opérations. La construction de variation sur ces problèmes permet d'étudier quels facteurs du problème

favorise ou défavorise la représentation de la relation mathématique permettant de réaliser le problème en une opération seulement (Hakem et al., 2011; Hakem, Sander, & Labat, 2005, 2005).

Le diagnostic est enregistré dans un assemblage de 18 colonnes regroupées en plusieurs groupes de colonnes (cf. Figure 13).

| Groupes | Intitulé | Exemple (symbolique) | Colonnes |
|---|---|---|---|
| 0 | Stratégie | | 1* |
| 1 | Resultat Intermédiaire dans le calcul en étapes | T1-P1 | 2<br>3<br>4<br>5 |
| 2 | Terme de la comparaison | T1-d | 6<br>7<br>8<br>9 |
| 3 | Calcul de la différence entre les variables homologues | T2-T1 ou P2-P1 | 10<br>11<br>12<br>13 |
| | | | 14* |
| 4 | Calcul final | (T1-d)+(P1-d) | 15<br>16<br>17<br>18 |

| Index | Intitulé | Exemple de valeurs (simlifiés) |
|---|---|---|
| a | Type de calcul | implicite, addition à trou, soustraction… |
| b | Pertinence des données de l'opération | données correctes, non identifiable |
| c | Exactitude du résultat | correct, petite erreur, grosse erreur |
| d | Formulation du résultat | cohérente avec l'opération, non cohérente … |

| Colonnes spéciales | Intitulé | Exemple de valeurs (simlifiés) |
|---|---|---|
| 1* | Stratégie | étapes, différence, étape-différence |
| 14* | Nature de ce qui est calculé | tout, partie, autre |

Figure 13. Colonnes de codage de diagnostic dans DIANE.

La première colonne dans la grille de codage concerne la stratégie. Elle indique de manière globale le choix de résolution (pouvant être mixte) que l'élève a pris.

Le reste des indicateurs sont représentés par des groupes de 4 colonnes.

- Les colonnes 2 à 6 concernent le calcul du résultat intermédiaire dans le cadre d'une stratégie de calcul à étapes.

- Les colonnes 6 à 9 concernent le calcul de la valeur du terme de la comparaison inconnue à partir de la valeur du terme connu et de la différence, elles concernent les cas où une comparaison est décrite dans l'énoncé.
- Les colonnes 10 à 13 concernent le calcul de la différence entre les variables homologues des deux parties de l'énoncé tout1 et tout2 ou partie1 et partie2. Elles retranscrivent les cas où le soldeur utilise une stratégie par différence.
- La colonne 14 concerne la nature de la solution finale calculée (partie, tout, comparaison, autre. ).
- Les colonnes 15 à 18 concernent le calcul de la réponse à la question finale.

Les 4 colonnes de chaque groupe suivent la même logique. La première caractérise un type de calcul (addition, soustraction, addition à trou, etc.), la seconde indique si les données de l'opération sont correctes, la troisième indique si le résultat est correct et la quatrième indique s'il y a une formulation du résultat, et si elle est correcte. Le diagnostic automatique est construit pour répondre à un besoin d'efficacité dans l'analyse de protocole de résolutions. En effet, en tant qu'activité manuelle, la tâche est fastidieuse et demande un niveau d'expertise élevé (Hakem, Sander, & Labat, 2005).

## 5.6.2 Limites de DIANE

### 5.6.2.1 Description des problèmes

Dans DIANE, l'interface de création de problèmes laisse la possibilité de renseigner quelques informations cruciales portant sur les énoncés produits (structure du problème et place de l'inconnue). D'autres, ne jouant pas un rôle particulier dans le diagnostic, sont aussi présents, avec l'idée qu'ils pourront jouer un rôle plus tard dans des diagnostics ou être utilisés dans le cadre d'une analyse de donnée comme la charge en mémoire de travail ou la nature de la variable dans les problèmes (cardinal, ordinal...). Cependant, la recherche étant par définition en mouvement, d'autres types d'informations sur les problèmes pourraient être envisagés. Or, dans la version actuelle, il faudrait, pour chaque nouvel indicateur, modifier à la fois l'interface et la base de donnée, ce qui est coûteux. Outre l'ajout d'indicateurs, si un nouveau type de problème mathématique est construit dans DIANE, le temps de développement peut-être très élevé, car il est requis de développer un nouveau script d'analyse réalisant le diagnostic comportemental du sujet sur ce type de problème. Pour que DIANE soit un

environnement favorable à la modélisation cognitive à long terme, il est important de pouvoir représenter les problèmes à un haut niveau de généricité.

Pour que DIANE soit utilisable sur le long terme dans une perspective de recherche interdisciplinaire, il est aussi nécessaire qu'il soit adopté par les professeurs. En effet, si un professeur a du mal à utiliser l'interface ou qu'il veut produire un problème dont la catégorie d'appartenance n'est pas implémentée dans DIANE, la déconvenue peut être importante. L'interface actuelle est en effet un peu lourde pour le professeur qui souhaite proposer un problème de sa création. Par ailleurs, s'il passe à côté de la logique de la typologie de problème proposée, le diagnostic ne peut pas fonctionner correctement. Le cas des problèmes complexes est encore plus sensible, car le professeur doit créer son énoncé en plusieurs blocs. L'ordre des informations est très important dans la résolution. S'il n'est pas respecté, alors le module de diagnostic interprète faussement les réponses des apprenants. De plus, les nombres de l'énoncé doivent être finement choisis pour éviter au module de diagnostic de rencontrer des cas ambigus. L'interface de création de problèmes représente donc un enjeu important pour généraliser l'emploi de DIANE dans les classes tout en facilitant son emploi et son évolution pour des études en psychologie cognitive.

#### 5.6.2.2 Diagnostic comportemental dans DIANE

Le diagnostic comportemental dans DIANE est la capacité principale qui différencie ce projet des différents outils visibles dans l'état de l'art. Si le système de création de problèmes est repensé, il est nécessaire que le module de diagnostic soit en accord avec ces évolutions. En effet, l'organisation par type de problème ne concerne pas seulement l'enregistrement dans la base de données des diagnostics, mais aussi la manière dont le diagnostic de DIANE est établi. Chaque type de problème a son propre script de diagnostic. De ce point de vue, il est conçu à un niveau de généralité extrêmement faible, car il n'y a pas de fonctions ou de classes communes à ces différents scripts. Cela pose un problème de maintenance, mais aussi d'évaluation scientifique. Un module de diagnostic plus modulaire, par le biais de la suppression ou la modification des modules qui le constituent et de l'observation des changements de performances associées peut renseigner sur la pertinence des critères utilisés dans la construction du module de diagnostic. Toutefois, avec des performances reportées dans (Hakem et al., 2011; Hakem, Sander, Labat, et al., 2005) suggérant un taux de succès dans le diagnostic comportemental supérieur à 90 %, ces scripts étaient efficaces, et nous donne

l'assurance que la méthode d'analyse consistant à éliminer le texte et se baser uniquement sur les expressions numériques est suffisante.

Un deuxième point important à soulever est que le module de diagnostic de DIANE porte sur un aspect limité des réponses de l'apprenant. En effet, DIANE a été fondée pour un projet de recherche visant l'identification de stratégies de résolution, de ce fait, la manière dont sont diagnostiquées les réponses dans l'outil est calibrée pour remplir cet objectif. Il en résulte donc que la grille de codage des réponses n'est pas purement descriptive, car elle est construite en référence aux différentes stratégies de résolutions des problèmes. Le diagnostic comportemental des problèmes complexes dans DIANE ne peut pas capter toutes les erreurs des sujets, il capte seulement les erreurs qui s'apparentent à différentes stratégies de résolutions préconçues. Cette remarque nous amène à la réflexion suivante : est-il réellement souhaitable de construire un diagnostic comportemental neutre ? L'avantage d'un tel diagnostic, qui chercherait seulement à décrire le chemin de résolution de l'apprenant est qu'il ne ferme pas la voie à d'autres analyses cognitives du comportement. C'est la notion de diagnostic comportemental tel que nous l'avons abordé dans notre partie théorique. Cependant, un diagnostic comportemental « neutre » est inutilisable en l'état. Il est nécessaire de l'enrichir pour faire des inférences qui ont un sens pédagogiquement ou dans le cadre d'une recherche expérimentale. Si nous cherchons à « généricicer » le diagnostic comportemental de DIANE, il reste crucial de laisser la possibilité à l'environnement de « reconnaître » des types de réponses particulières spécifiées par le chercheur.

### 5.6.3 Ouvertures pour le développement de DIANE

Dans cette partie, nous explorons certains travaux dont l'adaptation pourrait permettre à DIANE de résoudre les difficultés que nous avons identifiées. Les deux sous-parties présentées font respectivement écho aux deux limites représentées par les deux sous-parties plus haut.

#### 5.6.3.1 Conception et réutilisation de gabarits

La conception d'EIAH représente un travail important. Pour diminuer la difficulté de la tâche, de nombreux chercheurs ont développé des « outils auteurs », c'est à dire des outils qui facilitent le développement de systèmes et adaptés à un public non informaticien. Murray (2003) propose un état de l'art des outils auteurs dans les EIAH rassemblant de nombreux outils. Sept grandes catégories ont été construites par son travail de synthèse. DIANE peut-être assimilée à la classe « objectif spécifique »

(special purpose), c'est-à-dire les outils qui portent sur un domaine d'apprentissage assez précis. Dans cette catégorie, les fonctionnalités « auteurs » de ces outils sont le plus souvent portées par un système de gabarits. Un gabarit est un patron qui permet de fournir une préconstruction de ce que l'auteur veut produire, le « texte à trou » par exemple, est une forme de gabarit possible. Il suffit de modifier des paramètres dans le gabarit pour générer ce qui est visé (dans notre cas : des problèmes arithmétiques). Cette approche permet de créer rapidement de nouveaux problèmes en conservant la puissance du tuteur sur ces problèmes[24].

Hors du domaine des PAEV, et dans le domaine de l'algèbre, l'analyse des réponses ouvertes est plus classique (H. Chaachoua et al., 2005; Delozanne, Prévit, Grugeon-Allys, & Chenevotot-Quentin, 2010). En effet, le diagnostic comportemental dans ce domaine est crucial pour pouvoir ensuite, par exemple, identifier les règles correctes ou erronées de transformation d'expressions algébriques employées par l'élève. Dans ce domaine, le projet Pépite (Delozanne, Prévit, Grugeon, & Chenevotot, 2008; El-Kechaï et al., 2011) entretient une certaine proximité avec DIANE, car son but est de construire des diagnostics cognitifs sur des réponses ouvertes. Le projet Pépite, sous le nom de PépiGen (Delozanne et al., 2008), a par ailleurs fait le choix d'utiliser un système de gabarit pour étendre les problèmes dans son système. Son diagnostic est échelonné sur plusieurs niveaux. Le premier, nommé diagnostic local, consiste à évaluer la réponse de l'apprenant en la catégorisant selon des types d'erreurs. Ces erreurs attendues, propres à chaque exercice, sont **enrichies par un code** qui permet d'établir un diagnostic global et cognitif l'élève sous la forme de stéréotypes permettant de former des groupes d'élèves selon leurs compétences. Les professeurs ont accès au diagnostic local et global en langage naturel. Pour étendre ses fonctionnalités, un générateur d'exercice a été conçu à partir de plusieurs patrons. Ces derniers permettent de générer un exercice avec des données numériques obtenues aléatoirement, mais respectant des contraintes numériques. Les gabarits sont mis en place dans Pépite par l'expert informaticien, il

[24] En l'occurrence, pour le cas de DIANE, il s'agit plutôt de module de diagnostic que de tuteur, mais l'idée reste inchangée.

n'est pas possible pour un utilisateur de créer le patron d'un nouveau type de problème. Une exception est cependant notable, ce sont les problèmes dits « du magicien », visibles en Figure 14. Pour ces problèmes spécifiques, le professeur peut construire son énoncé à l'aide de boutons cliquables. Le principe de construction de problèmes par le biais d'une interface composée de boutons semble a priori applicable aux PAEV.

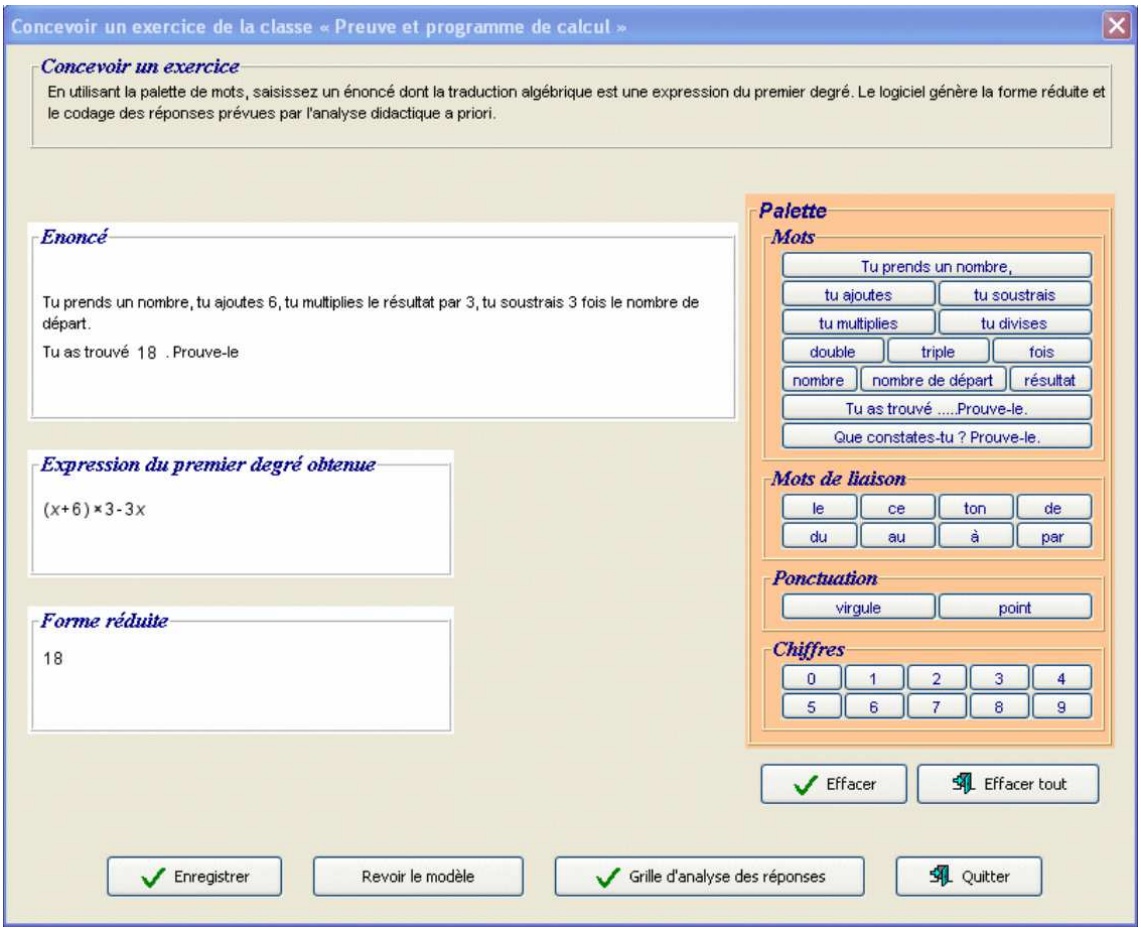



Figure 14. Création assistée d'un problème à énoncé verbal dans Pépite.

Les choix de conception sont intéressants et semblent être applicables au projet DIANE. Enfin, utiliser des patrons d'exercice semble être une démarche pertinente pour étendre le nombre d'exercices dans DIANE tout en préservant la comparabilité des performances à différents problèmes. Le principe de propriétés associées aux réponses semble aussi opportun, car il permet de stocker des informations relatives à ces observations potentielles. Dans le cadre de la trajectoire de généricité visant une approche flexible de la création de problèmes, il serait pertinent de laisser la possibilité au concepteur d'exercices de définir lui même ces propriétés enrichissant les problèmes et les réponses attendues. Si cette fonction d'étiquetage est ouverte, alors elle peut

permettre à plusieurs modèles, éventuellement concurrents, de s'exprimer sur les réponses dans DIANE.

La technique d'utilisation de gabarit devrait non seulement permettre de générer des instances d'exercices, mais aussi d'être modifiable et concevable par le chercheur. Il serait donc pertinent de séparer les droits administrateurs des droits chercheurs afin de permettre l'évolution des fonctionnalités laissant la possibilité au système d'être modifié pour répondre à de nouvelles questions de recherche ou à des problématiques de modélisation.

#### 5.6.3.2 Deep Path Finding

Ohlsson et Langley (1990) ont proposé une méthode pour automatiser le diagnostic comportemental dans les EIAH : DPF (Deep Path Finding). Comme nous l'avons entrevu dans la partie théorique, l'enjeu du diagnostic comportemental est de reconstruire, avec des données parfois lacunaires, le cheminement de l'apprenant. Ohlsson et Langley proposent de se placer dans le cadre de l'espace de recherche tel que Newell et Simon l'ont formulé, c'est-à-dire un graphe constitué d'états du problème et des opérations permettant de passer d'un état à l'autre. Le but est de trouver le chemin le plus plausible dans cet espace allant de la situation initiale du problème à l'état final qui correspond à la réponse formulée par l'apprenant. Le cas d'étude pris par les auteurs est le diagnostic comportemental des soustractions posées. Dans ce cadre, seule la réponse finale est connue du système. De ce point de vue, cette situation est analogue au diagnostic comportemental de réponses ouvertes dans les PAEV. La technique la plus simple pour réaliser un diagnostic comportemental dans ce cadre est l'emploi de « force brute », c'est à dire tester toutes les possibilités et prendre la première qui permette de reconstruire un chemin de résolution aboutissant à la réponse du sujet. Cette approche ne permet pas, toutefois, de sélectionner les chemins qui sont les plus plausibles d'un point de vue psychologique. Ohlsson et Langley listent donc une série de principes qu'ils ont choisi d'implémenter pour permettre une sélection intelligente des chemins de résolution concurrents :

- Causal Closure : seuls les états directement accessibles par le biais d'opérateurs et de données du problème (ou obtenues en cours de résolution) sont considérés.
- Purposefulness : les cas dans lequel la production d'une opération est réutilisée plus tard sont favorisés. Les auteurs notent que cette clause est parfois fausse,

mais toutes choses étant égales, les chemins respectant cette propriété doivent être privilégiés.

- No duplication : Les sujets ne recalculent pas des données ou des résultats intermédiaires déjà accessibles.
- Memory Load : La capacité en mémoire de travail des humains est limitée, donc les chemins qui emploient un grand nombre de résultats intermédiaires non utilisés sont défavorisés.
- Subgoaling & productivity : Les chemins qui remplissent le plus de sous-buts possibles et avec le moins d'opérateurs possible sont favorisés.
- Minimal Error : Les chemins qui sont les plus proches d'une bonne solution sont favorisés.
- Minimal Length : Les chemins qui utilisent le plus faible nombre d'opérateurs sont favorisés.

DPF constitue donc un cas d'étude intéressant dans la mesure où la construction d'un module de diagnostic amélioré pourrait implémenter la plupart de ces critères.

# 6 PROBLEMATIQUE

La cause que nous avons considérée comme la plus probable du manque de profondeur du diagnostic dans les EIAH est la recherche d'une amélioration de l'apprentissage avec des coûts limités. Le développement de système de diagnostic, mettant, au moins temporairement, l'enjeu du tutorat de côté, est source d'opportunités pour le développement d'approches cognitives profondes dans la modélisation de l'apprenant. Ils permettraient à des modèles cognitifs plus détaillés d'être développés en analysant le comportement humain par la cohérence de ses erreurs. Cependant, dans le domaine des PAEV, les programmes évitent généralement les complications en se cantonnant à une modélisation du comportement limitée à l'échec ou la réussite d'un problème.

La première étape dans la réalisation d'un tel système de diagnostic est de pouvoir former un diagnostic comportemental qui puisse à partir d'une réponse brute d'un élève sur un PAEV, effectuer son diagnostic comportemental pour qu'elle soit intelligible par le système. Dans des démarches ambitieuses de diagnostic cognitif, il est nécessaire d'avoir un diagnostic comportemental riche. La construction de ce type de diagnostic est donc un pas important. Comme nous l'avons vu, le diagnostic comportemental est une problématique classique dans les EIAHs. Cependant, sa richesse de ce diagnostic est contrainte par son emploi ultérieur. À de rares exceptions près, dans l'ensemble des EIAHs portant sur les PAEV, le diagnostic comportemental est limité. Très peu de systèmes dédiés aux diagnostics ont été développés, et, hormis le projet DIANE, aucun programme de diagnostic ne proposait une analyse des réponses allant au-delà de la détection de justesse ou fausseté de la réponse. Plusieurs améliorations sont pourtant souhaitables pour le diagnostic de DIANE. Les scripts d'analyse de DIANE manquent

de transparence et ne sont pas suffisamment modulaires pour les décrire et les évaluer de manière précise. Par ailleurs son diagnostic reste construit sur la référence des stratégies de résolutions possibles aux différents problèmes. Une reconstruction plus ambitieuse du système de diagnostic de DIANE est donc visée pour répondre aux exigences présentées. Si des recommandations générales existent (Langley et al., 1990) pour l'établissement de diagnostic, leurs adaptations aux PAEV restent à faire. Une problématique supplémentaire, proche à la dynamique transdisciplinaire que nous souhaitons mettre en place est l'évaluation de la fiabilité du diagnostic produit. Pour soutenir des recherches en psychologie cognitive, il est important de savoir quel niveau de confiance il est possible d'attribuer au module de diagnostic. Pour diagnostiquer le module de diagnostic lui-même, une démarche expérimentale est souhaitable. Nos questions sont donc les suivantes : **comment mettre en place un module de diagnostic adapté aux PAEV, transparent, générique ? Comment évaluer la qualité générale du programme et mesurer la fiabilité de ses diagnostics ?**

Dans une démarche de modélisation cognitive, l'analyse des erreurs est cruciale pour étudier les représentations et les connaissances du sujet. Pour pouvoir construire un système optimisé pour l'arithmétique en classe élémentaire, il est souhaitable de disposer d'un éclairage de la littérature sur le domaine. Dans le chapitre analysant les sources d'erreurs dans la résolution de PEAV, nous avons analysé plusieurs approches conflictuelles. L'emploi des mots-clefs dans le cadre de stratégies superficielles est un phénomène souvent suggéré dans la littérature. D'autres approches contrastent cette vision, celle des difficultés linguistiques et des réinterprétations et le respect de contraintes didactiques qui contraignent la résolution. Prises seulement en tant que théories générales, ces approches ont peu d'utilité dans le cadre de systèmes de diagnostic. Pour pouvoir être employées dans un tel environnement, il est important d'être capable de les mettre en lien avec les productions des apprenants. D'un point de vue cognitif, intégrer ces théories dans un modèle « computationnel » (exécutable), permettrait d'évaluer plus quantitativement leurs capacités explicatives des erreurs des élèves respectives. Un modèle permet aussi de les placer en concurrence et d'évaluer leurs complémentarités, ce qui est, par définition, impossible dans des études compartimentées. Ce type de modèle se fait rare dans la littérature. Dans la construction de notre modèle, nous nous posons donc la question générale suivante : **quels sont les pouvoirs prédictifs de ces différentes approches ?** De manière préliminaire, la question des méthodes statistiques de validation doit être posée. Nous avons pu voir lors

de l'analyse de Sierra que l'analyse des modèles génératifs peut être problématique du fait que les statistiques associées descriptives et non inférentielles. **Comment, dans le cadre des PAEV, quantifier la qualité des prédictions dans un cadre conforme aux méthodes statistiques usuelles en psychologie cognitive comme les tests d'hypothèses ?**

Si cette étape concernant l'association des erreurs à leurs sources hypothétique est réalisée, d'autres étapes sont requises pour former un système de diagnostic cognitif qui réponde à nos exigences. Deux grandes catégories de modèles existent : les modèles numériques et les modèles symboliques. Nous avons pointé que les méthodologies de diagnostic des conceptions erronées dans les approches numériques pouvaient aussi être critiquées par sa non-utilisation des prédictions précises des règles erronées ou bien par explosion du nombre de paramètres libres. L'approche réussite/échec est suffisante pour les modèles d'attribution de compétences, mais peut être insuffisante dans le cadre d'une modélisation cognitive plus profonde. En effet, s'il est souhaité modéliser la résolution de PAEV de manière cognitive, il est préférable de travailler à un niveau plus détaillé pour faire intervenir des connaissances ou hypothèses sur les processus de résolution et les possibles conceptions erronées. Par ailleurs, la recherche de traits latents n'est pas une approche profonde du point de vue de la psychologie cognitive qui préfère la notion de processus. Dans un contexte de modélisation cognitive, les modèles symboliques sont donc à la fois une source d'opportunités et présentent d'importants obstacles. Leur caractère déterministe, cependant, pose problème. Nous avons montré l'existence d'une tension de ces modèles avec les modèles numériques, synthétisé dans le commentaire des éditeurs du livre Cognitively Diagnostic Assessment. Ces modèles sont à la fois source d'opportunité pour former des modèles de grande qualité, mais gênants dans leur gestion de l'incertitude et de leur validation. Les modèles symboliques, dont le modèle des contraintes fait partie, permettent potentiellement de formaliser le comportement avec une bonne pertinence psychologique, mais sont largement minoritaires dans les systèmes de diagnostic contre des approches psychométriques plus faciles à mettre en place et dont l'évaluation statistique de la pertinence est très bien balisée. En revenant à la question fondamentale de la validation statistique des modèles qui est la mesure sa qualité tout en prenant en compte le nombre de paramètres libres employés à sa construction, nous avons listé des critères habituellement utilisés dans le cadre des modèles probabilistes. Un principe, le MDL, a retenu notre attention et a le potentiel pour être applicable pour mesurer la qualité des

modèles déterministes sans les transformer en modèles probabilistes. Outre l'implémentation d'un calcul du MDL adapté, il est souhaitable de définir un espace d'écriture de modèles qui puissent être mesurés avec le critère constitué. Au travers de notre état de l'art, nous avons souligné que le problème du temps de développement des modèles cognitifs était problématique. Le temps de développement est un argument utilisé en faveur des Constraint-Based Tutors contre les Cognitive Tutors. Richard pointe que la mise en place du modèle cognitif utilisé dans la modélisation de la correction de défaut dans la manipulation de presse à injecter est coûteuse. De nombreux outils auteurs (Murray, 2003) ont été développés pour contourner cet obstacle. Ignorer la contrainte du coût pose un risque, celui de développer des approches qui ne seront pas réemployées dans le cadre d'EIAH, faute de moyens. Il est donc souhaitable de la prendre en compte dans notre approche. Outre la mise en place d'outils, notre thèse se centre sur les PAEV. Il sera donc alors question de réemployer les résultats obtenus dans la modélisation des erreurs pour les faire intervenir à un niveau interindividuel, dans l'établissement de diagnostics cognitifs différentiels. Ainsi, en troisième partie, nous nous poserons les deux questions suivantes : **comment mesurer la qualité d'un modèle déterministe ? Comment faciliter le processus itératif de construction et de validation de ces modèles ? Comment réaliser un diagnostic cognitif de la résolution de PAEV ?**

## Données utilisées pour valider les modèles

Nos travaux sont centrés sur la modélisation, le coût en temps du développement de ces approches et la quantité de données nécessaires à la validation de modèles nous ont amenés, dans la première et la deuxième partie, à utiliser des données issues d'une autre étude. Nous avons choisi de tester notre modèle avec un ensemble de données externes pour pouvoir nous concentrer sur la modélisation et nous dégager des contraintes de l'expérimentation. Nous avons pu obtenir des données des deux expérimentations sur la résolution de PAEV complexes (Chaillet, 2014). Ces données sont codées à un niveau d'analyse suffisamment fin pour tester nos modèles.

Le matériel est composé de deux groupes de 16 problèmes que nous appelons « problèmes complexes » dont nous présentons deux exemples :

Tc4t :

> *Au supermarché, le kilo de poisson a augmenté de 5 euros cette année. Un kilo de poisson coûte maintenant 12 euros. Au début de l'année, le kilo*

*de viande coûtait le même prix que le kilo de poisson. Le kilo de viande a augmenté de 3 euros de moins que le kilo de poisson. Combien coûte le kilo de viande maintenant ?*

Cc1p :

*Antoine (P1) a 5 billes. Quand Antoine (P1) réunit ses billes avec celles de Paul(P2), ils ont 12 billes ensemble(T1). Quand Paul(P2) et Jacques(P3) réunissent leurs billes, cela fait 3 billes de moins (d). Combien Jacques a-t-il de billes ?*

Nous nous référons à ces problèmes et aux quantités qu'ils mettent en jeux de manière codifiée qu'il s'agit d'expliciter. C'est la raison pour laquelle nous avons rajouté des lettres devant les quantités. Elles prennent place dans des schémas appelés partie-tout. Dans le deuxième exemple, les billes d'Antoine et les billes de Paul sont les deux parties d'un tout. Nous utilisons la lettre majuscule **P** pour désigner une partie et **T** pour désigner un tout. Chaque problème décrit deux touts (**T1** et **T2**), le premier est composé des parties **P1** et **P2**, le deuxième est composé des parties **P2** et **P3**. Le fait que ces deux schémas ont une partie commune offre la possibilité de multiples stratégies de résolution (cf. Figure 15). Tous ces problèmes peuvent être résolus en 1 ou 3 étapes. En revanche, la quantité demandée dans la question posée à la fin de l'énoncé est différente dans les deux groupes. Elle concerne soit la partie **P3** soit le tout **T2**. Au sein de chaque groupe, les formulations et les contextes sémantiques varient, c'est pour cela que leur nom de code change. Les 16 problèmes sont écrits en annexe.

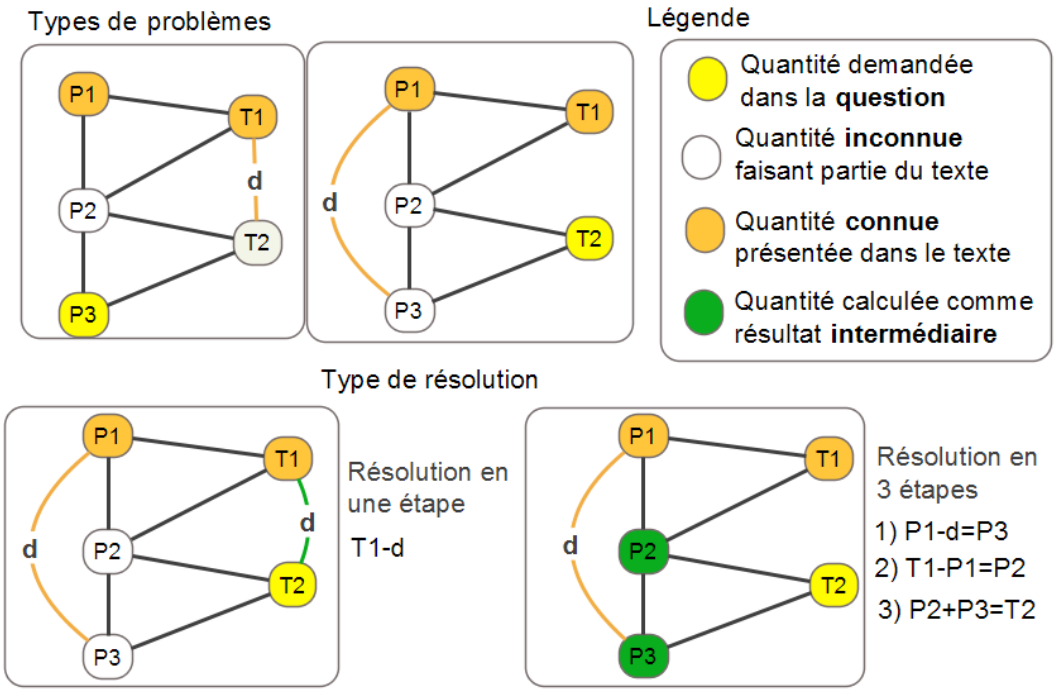



Figure 15. Représentation des deux structures mathématiques des problèmes étudiés. En orange, sont représentées les trois quantités données dans l'énoncé, en jaune la quantité à trouver. Les liens indiquent des relations mathématiques entre les quantités.

Les problèmes se décomposent en trois propriétés : le contexte sémantique (effectif, âge, poids, argent), le type de question (partie ou tout) et le type de relation de la première partie du problème (transformation ou complément). Ainsi Tc4t signifie « problème de transformation (**T**) , avec le contexte argent(**4**), question sur le tout (**t**) »

La première rassemble 388 enfants de CM1-CM2 (203 élèves de CM1 et 185 élèves de CM2) ; l'âge moyen des enfants est de 10 ans et 1 mois avec un écart-type de 10 mois. Les classes se situent dans Paris, la région parisienne ainsi qu'en province, et le cadre socio-économique est diversifié. Les enfants devaient résoudre 12 problèmes, dont 8 problèmes expérimentaux. La seconde rassemble 384 enfants de CM1-CM2 (192 de chaque classe) à Paris et en région parisienne ; la situation socio-économique est également diversifiée. L'âge moyen des enfants est de 10 ans et 9 mois avec un écart-type de 9 mois. Pour cette seconde expérimentation, les enfants devaient résoudre 7 problèmes, dont 4 problèmes expérimentaux. Nous avons au total pu obtenir 6176 réponses, dont 2854 « mauvaises réponses ».

Le but de ces recherches était d'étudier l'influence de ces variables sur les performances et les stratégies de résolution. Il est notamment montré que les problèmes ayant une question portant sur le tout sont mieux réussis que ceux portant sur la partie. Les

problèmes qui sont du type complément sont aussi plus facilement réussis que ceux de transformation. Les stratégies de résolutions varient selon le contexte sémantique et le type de question posée. La deuxième expérience, qui porte, comme sur la première, sur des CM1 et des CM2 avaient plusieurs buts : confirmer les résultats de la première expérimentation en rectifiant certains éléments des problèmes de la première expérimentation et par le biais d'un autre dispositif expérimental, demander aux enfants de deviner la question de certains problèmes puis d'y répondre ensuite. Cette étude répondait à une autre question théorique que nous résumons très brièvement ainsi : est-ce que les enfants se trompent, car ils infèrent les questions sans les lire ?

Les données que nous présentons représentent des qualités intéressantes pour nos analyses visant à diagnostiquer les erreurs :

- La base de données est large et contient beaucoup d'erreurs (la moitié des réponses environ). La variété des réponses est un atout important en ce qui concerne nos projets de modélisation. Le fait que ces problèmes comportent trois nombres offre un espace de recherche assez large pour établir des prédictions et va donc permettre à un modèle cognitif de s'exprimer librement dans le deuxième chapitre de nos contributions.

# DEUXIEME PARTIE

# CONTRIBUTIONS

# 7 MODELISATION COMPORTEMENTALE. COMPRENDRE LA REPONSE DE L'APPRENANT.

## 7.1 Nouvelle version de l'interface de création de problèmes

Nous avons soulevé dans les parties précédentes qu'un mode de création de problèmes plus ouvert est souhaitable pour accompagner une dynamique de recherche en psychologie cognitive au sein de DIANE. Nous avons par ailleurs identifié que les systèmes à base de gabarit pouvaient répondre à ce besoin. Ce travail était important pour travailler dans un cadre plus adéquat dans nos travaux de modélisation. Pour ne pas alourdir le document, nous plaçons la description de nos travaux en annexe et nous ne donnons qu'un résumé dans cette partie.

La création d'un problème dans DIANE permet actuellement de définir des propriétés sur les problèmes. Ces propriétés agissent comme des « étiquettes » et le chercheur a accès à des interfaces lui permettant d'ajouter, supprimer des propriétés ainsi que de les associer à des problèmes nouveaux ou déjà crées. Cet ajout de propriété permet une « indexation cognitive » des problèmes qui peut évoluer au cours du temps. Symétriquement, il est maintenant possible d'ajouter à chaque problème des réponses attendues auxquelles il est possible d'ajouter, tout comme pour les problèmes, des propriétés. Les propriétés sur les réponses attendues sont aussi ouvertes que les

propriétés sur les problèmes. Ces fonctionnalités ne sont ouvertes qu'à l'expert du domaine. Pour ouvrir l'outil aux professeurs, une fonctionnalité de « clonage » a été mise en place, elle permet au professeur des écoles de récupérer la richesse de l'outil auteur (produire des problèmes nouveaux) sans la complexité de la création de problème (clonage réalisé en un clic) et sans altérer le problème du concepteur (propriétés mathématiques conservées).

Ce jeu de descripteurs ouvert, tout comme la formation d'un nouveau module de diagnostic comportemental qui fait l'objet de la partie qui suit, laisse des possibilités d'analyses nouvelles. Dans le dernier chapitre de nos contributions, les propriétés des problèmes ainsi que les propriétés sur les réponses fournies par les réponses des élèves sont exportées et analysées par le programme que nous décrirons.

## 7.2 Implémentation d'un diagnostic comportemental générique

Le but d'un diagnostic comportemental générique est d'identifier les calculs établis par l'élève. Le diagnostic comportemental, comme dans les approches précédentes (Hakem et al., 2011; Hakem, Sander, Labat, et al., 2005), se base uniquement sur les formules mathématiques et sur les nombres donnés dans la réponse de l'élève. L'analyse de la réponse est produite par le travail conjoint de quatre différents modules que nous décrivons ci-après.

- Un module de prétraitement s'assurant que la réponse est sous une forme qui peut être traitée par la suite (**PT**).
- Un premier module d'analyse globale (**AG_1**).
- Un module d'analyse séquentielle (**AS**) qui interprète tour à tour les formules de l'élève. Il est lui-même composé de sous-modules.
- Un deuxième module d'analyse globale (**AG_2**) qui établit les derniers éléments du diagnostic.

### 7.2.1 Module de prétraitement (**PT**).

La première étape consiste à effectuer un prétraitement de la réponse des élèves. Il s'agit, si elle présente certaines difficultés pour les scripts suivants, de la remplacer par une formule équivalente qui ne produira pas d'échec. Comme le diagnostic se base sur les nombres, nous remplaçons les termes contenant un chiffre (comme « CM1 », « CE2 »...) par des termes sans chiffres tels que « CM_a », « CE_b », etc. Étant donné

que dans le module suivant, les nombres en toutes lettres sont remplacés par leur équivalent numérique (exemple : « deux » devient 2), certaines expressions sont transformées pour éviter les faux positifs (exemple : l'expression « tous les deux » devient « ensemble »). Enfin, des équivalents sont utilisés pour certaines formules inhabituelles afin d'éviter de mettre en échec les modules qui les traitent. Trois types de réparation sont effectués :

- Les formules du type a+b=c-d=e. Ces formules résultent d'une conception erronée du signe égal (Kieran, 1981). Le signe égal ne représente pas, pour beaucoup d'enfants, la notion d'équivalence, mais est plutôt vu comme un opérateur. L'opération écrite pourrait se traduire à l'oral par *« 3 plus 8 cela fait 11, moins 4 cela fait 7 »*. Par principe d'économie, 11 n'est pas représenté deux fois, et l'interlocuteur comprend le calcul transmis.

- Les formules du type a+b-d=e. Contrairement au type de formule précédente cette forme n'est pas erronée, mais la traiter différemment des autres nuit à la généricité du module suivant permettant de traiter les formules du type « a+/- b = c ». Elles sont donc détectées et remplacées par des équivalents du type a+b=c , c-d=e.

- Le dernier type de formules modifiées concerne celles sans signe égal. Un exemple est 3+8 11. Pour éviter les risques de faux positifs, elles sont détectées seulement si deux conditions sont réunies : (1) une formule incomplète apparait « 3+8 », (2) le nombre suivant est le résultat exact de l'opération précédente.

### 7.2.2 Module d'analyse séquentielle **(AS)**

Nous ne détaillons pas le premier module d'analyse globale **AG_1**, celui-ci relève quelques indicateurs et liste les formules (e.g : 3+6=9) dans la réponse de l'apprenant. Cette liste de formules va ensuite être traitée par le module que nous étudions ici (**AS**), ce module est constitué des opérations les plus complexes dans l'établissement du diagnostic. Nous reprenons des critères mis en avant par DPF développé par Ohlsson et Langley (1990) pour guider nos choix lorsque nous devons sélectionner l'hypothèse la plus plausible psychologiquement lorsqu'une ambigüité se présente.

#### 7.2.2.1 Informations accessibles au module

Ce module utilise et maintient à jour un **espace de travail global** qui liste l'ensemble des informations accessibles à l'enfant à chaque étape de sa résolution. Ce travail

d'espace global permet de mettre en place le critère que Ohlsson et Langley (1990) nomment « Causal Clausure » qui consiste à ne considérer que les états accessibles à partir des informations courantes. Cet espace est constitué de trois listes de nombres à jour :

- Les nombres de l'énoncé (l1)
- Les nombres calculés au cours de la résolution (l2)
- Les nombres accessibles mentalement (par le biais d'un seul calcul impliquant les nombres de l'énoncé et les nombres calculés) (l3, déduite de l1 et l2)

L'espace de travail global contient donc les informations cognitivement accessibles à l'apprenant. Nous notons que la troisième liste contient des nombres qui ne sont pas strictement accessibles au même titre que les nombres dans l1 et l2. Nous l'utilisons pour essayer de comprendre chacune des formules rencontrées. Parfois, ces informations sont insuffisantes. Le module va inspecter chacune des formules. Tout nombre n'apparaissant pas dans les listes l1 et l2 a donc un statut spécial dit « inconnu », il peut être (ou ne pas être) dans l3.

#### 7.2.2.2 Fonctionnement du module

Pour qu'une formule puisse être interprétée, il faut pouvoir identifier le résultat dans la formule. En effet, celui-ci n'est pas toujours à droite du signe « égal ». L'addition à trou, par exemple, a pour nombre calculé le nombre situé en deuxième opérande (x+?=z). La condition, donc, pour distinguer le résultat des opérandes est de trouver quel nombre est « nouveau » dans la formule. Or, cette déduction n'est pas toujours facile. La méthode que nous proposons consiste à commencer par compter le nombre de chiffres inconnus.

Deux cas sont possibles :

- Le nombre d'inconnues dans la formule est égal à 1. Elle est non ambigüe, car l'inconnue est simplement le résultat de la formule écrite par l'élève. Ce cas est le plus simple. La formule est alors traitée par le module dédié qui décrit le type, la forme de l'opération et la présence éventuelle d'erreur de calcul. Par ailleurs, il corrige les erreurs d'écritures d'opérande tel « 12+8=4 » suggérant que l'élève, même s'il a utilisé le signe « + », a effectué une soustraction. C'est une fonction de réparation semblable aux fonctions vues dans le premier module. Pour qu'elle soit utilisée, le résultat doit devenir exact en changeant le signe.

Enfin, ce module informe celui de traitement séquentiel qu'un nouveau nombre est calculé. Ce dernier met donc à jour ses listes l2 et l3.

- Le nombre d'inconnues est supérieur à un. Nous supposons qu'un ou plusieurs nombres sont issus d'un calcul mental non explicité par l'élève. Nous regardons alors combien d'inconnus appartiennent à la liste l3, ce qui équivaut à chercher si un calcul mental peut expliquer ce nombre. Pour qu'il soit pris en compte, il ne doit pas comporter de faute de calcul. Trois sous-cas sont alors possibles :
    - Aucun calcul mental ne peut expliquer la présence des inconnues supplémentaires. Il n'existe pas de technique fiable pour désambigüiser ce calcul. La formule est considérée comme ininterprétable et est ignorée par la suite.
    - Des calculs mentaux sont accessibles pour désambigüiser le calcul, sans saturer la formule, c'est-à-dire qu'il reste **un seul nombre inexpliqué**, c'est donc le résultat de l'opération écrite par l'élève. Le calcul mental (ou les deux calculs mentaux si tous les nombres de la formule étaient inconnus) est alors ajouté dans la liste des formules effectuées et donc renvoyé au module de traitement de formule comme si l'élève l'avait écrit dans sa solution.
    - La formule est saturée : tous les nombres sont explicables comme le résultat de calculs mentaux. La stratégie employée est d'écarter le calcul mental candidat jugé le moins probable. Nous expliquons plus loin comment cette décision est prise dans la partie « Stratégies d'arbitrage dans les opérations sensibles ».

Le fonctionnement de ce module d'analyse séquentiel est présenté en Figure 16. Elle présente aussi les opérations de désambigüisation de formules que nous traiterons plus loin.

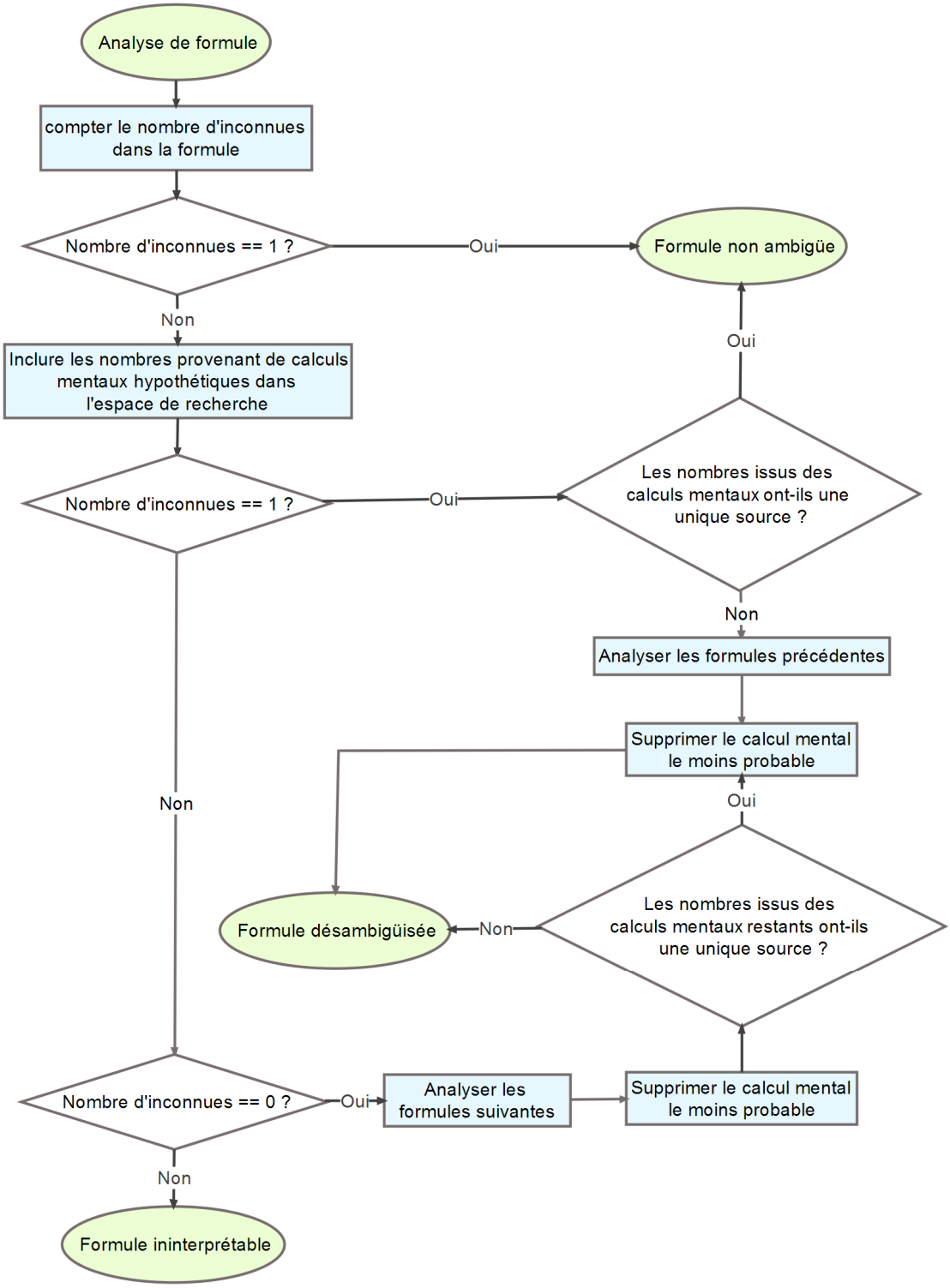



Figure 16. Synthèse du module d'analyse de formules

Le sous-module traitant les formules non ambigües prend une forme prend en entrée une formule qui est considérée comme réalisée par l'élève (non ambigüe ou désambigüisée) et extrait les informations suivantes.

- Type d'opération

  - Addition
  - Soustraction
- Forme d'opération
  - Addition simple
  - Addition à trou
  - Soustraction simple
  - Soustraction à trou
  - Soustraction inversée (3-8=5, se lit « 3 pour aller à 8 font 5 ».)
  - Opération mentale
- Présence éventuelle de fautes de calcul.
- Formule symbolique (T1-d). Si le calcul s'appuie sur des nombres calculés précédemment, elle peut être sous une forme composée telle que T1+(P1-d).

### 7.2.3 Deuxième module d'analyse globale (**AG_2**)

Le deuxième diagnostic global remplit de multiples objectifs. Deux cas peuvent se produire :

- La réponse de l'apprenant a été détectée dans la première partie « analyse globale », c'est-à-dire qu'un nombre isolé est donné à la fin. Il s'agit alors de l'interpréter au regard des formules de calculs qui ont été analysées dans le traitement séquentiel. Deux sous-cas sont possibles :
  - La réponse se trouve être le résultat d'un calcul de l'apprenant déjà identifié dans le module précédent. Elle est donc considérée comme telle.
  - La réponse de l'apprenant n'est pas le résultat d'une opération. Le programme regarde alors si elle peut être produite par un calcul mental. La méthode de détection de calculs mentaux définie précédemment est réutilisée pour effectuer cette analyse.
- La réponse de l'apprenant n'a pas été détectée dans la première analyse globale, il est alors considéré qu'elle est le résultat du dernier calcul proposé par le sujet.

### 7.2.4 Stratégies d'arbitrage dans les opérations sensibles

Dans la description du module, nous avons volontairement occulté deux opérations sensibles. Il s'agit dans cette partie de décrire le problème et les solutions abordées. Nous nous appuyons sur le fait que les protocoles de résolutions se comprennent dans une architecture de buts. Ainsi, comme nous l'avons vu dans la partie théorique, Deep Path Finding (Langley et al., 1990) utilise des critères utilisant le fait que le sujet effectue des opérations les unes après les autres avec des buts en tête ce qui implique, dans notre cas, qu'un nombre calculé a de bonnes chances d'être réutilisé plus tard. Nous nous basons sur ce critère pour désambigüiser les formules difficiles. Les deux problèmes peuvent être rencontrés dans la même réponse à diagnostiquer. Cet exemple nous permettra de les préciser : supposons que les nombres de l'énoncé soient les suivants : 14, 9, 2. Supposons ensuite que l'élève fait un premier calcul, explicite, « 14-9=5 » suivi d'un deuxième calcul, « 16-9=7 ». Il précise ensuite « la réponse est 16 ». Comme la Figure 17 le montre, interpréter la première formule ne pose pas de problème. La deuxième formule représente, elle, une difficulté, car deux inconnues sont présentes.

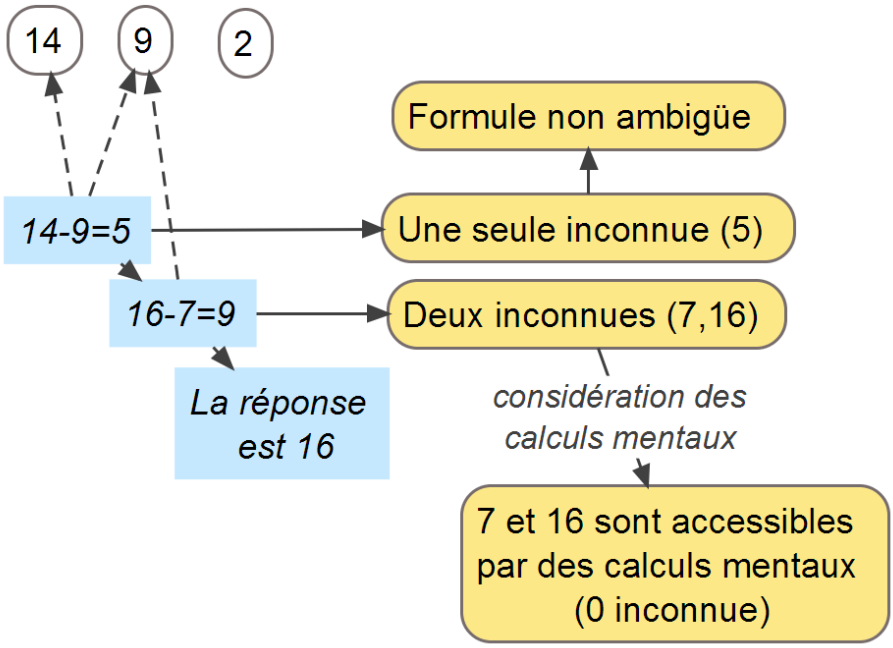



Figure 17. Explicitation des techniques de désambigüisation. Les cadres bleu clair représentent la réponse du sujet, les ronds représentent les nombres de l'énoncé. Les rectangles arrondis représentent les traitements et analyses effectuées.

Lorsque ce cas se présente, la liste des nombres accessibles par des calculs mentaux est considérée. Or, ici les deux nombres 7 et 16 peuvent provenir d'un calcul mental (14+2=16, ainsi que 9-2=7), c'est le cas que nous traitons ci-dessous.

**Problème 1 : Tous les nombres d'une formule ont une source possible**

Lorsque l'inspection de la possibilité qu'un opérande du calcul à désambigüiser provient de calculs mentaux, alors il peut arriver que tous les nombres du calcul résultent d'un calcul mental. C'est le dernier sous-cas que nous avons mentionné lors de la présentation du module d'analyse séquentiel. Ce phénomène est gênant, car il n'est pas possible de distinguer le nombre calculé des opérandes. Pour prendre une décision, nous sortons de l'espace de travail global pour inspecter ce qui pourrait être nommé « espace de travail futur » constitué des informations utilisées plus tard par l'élève dans sa résolution. Trois cas sont possibles, par **ordre de priorité** :

- Un unique nombre de cette équation est présent de manière isolée après la formule (compris alors comme la présentation d'un résultat).
- Si un nombre est utilisé dans la formule qui suit, il est supposé être le résultat calculé.
- Si aucun nombre n'est réutilisé plus tard, le nombre à droite du signe égal dans la formule à désambigüiser est considéré comme celui qui a été calculé.

Si nous revenons à notre exemple, le nombre 16 est effectivement réutilisé de manière isolé par la suite. Nous considérons alors que la formule a pour nombre calculé 16 et pour opérande 7 et 9 (cf. Figure 18). Conformément à notre ordre de priorité, si 16 n'était pas utilisé plus tard, alors c'est le 7 qui aurait été considéré comme « calculé »

car situé à droite du signe égal.

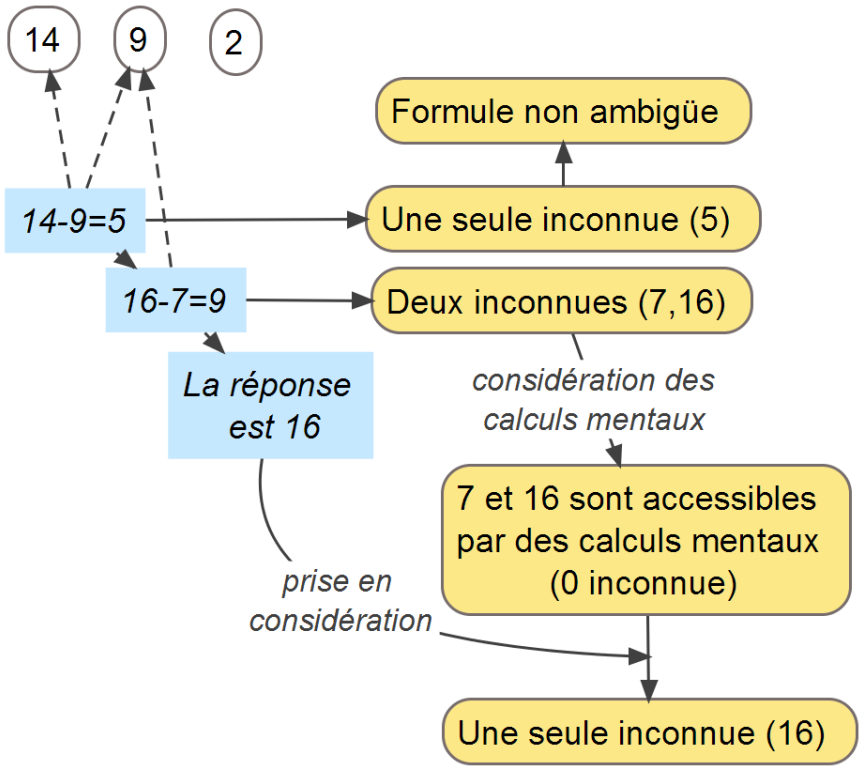



Figure 18. Considération des informations utilisées plus tard par l'élève pour désambigüiser une formule.

En réalité la réponse du sujet n'est pas encore totalement désambigüisée. Le problème est que 7 peut être issu de deux calculs mentaux différents ( 9-2 = 7 ou bien 5+2 = 7).

**Problème 2 : Un opérande a plusieurs sources possibles**

Plusieurs facteurs pour guider ces choix ont été pensés, mais pour implémenter un système de décision, le poids relatif de ces différents facteurs doit être pris en compte. C'est un obstacle important, car il introduit beaucoup d'incertitudes et de subjectivité dans le modèle de diagnostic. Pour cette raison, nous avons choisi de considérer que ce système, que nous appelons « politique de décision », ou plus simplement « politique », soit un paramètre modifiable du programme. La liste que nous avons constituée en tant que valeur par défaut est la suivante :

i. Dernier Nombre Calculé
ii. Nombre calculé
iii. Nombre situé après le dernier signe « égal »
iv. Nombre de l'énoncé

Cette liste est construite avec l'idée que l'enfant a plus de chance d'utiliser des nombres calculés récemment. Lorsque plusieurs sources expliquant la présence d'un nombre sont en compétition, cette liste est utilisée pour attribuer une note à chaque source. Lorsqu'il s'agit de décider quel calcul mental est le plus probable, la technique employée pour résoudre ce cas est la sommation des rangs dans cette liste. Dans notre exemple précédent, le calcul mental expliquant la présence du « 7 » serait donc celui qui s'appuie sur le dernier nombre calculé ainsi qu'un nombre de l'énoncé qui correspond au calcul numérique « 5+2=7 ». En effet, l'explication concurrente utilisait quant à elle deux nombres de l'énoncé (cf. Figure 19).

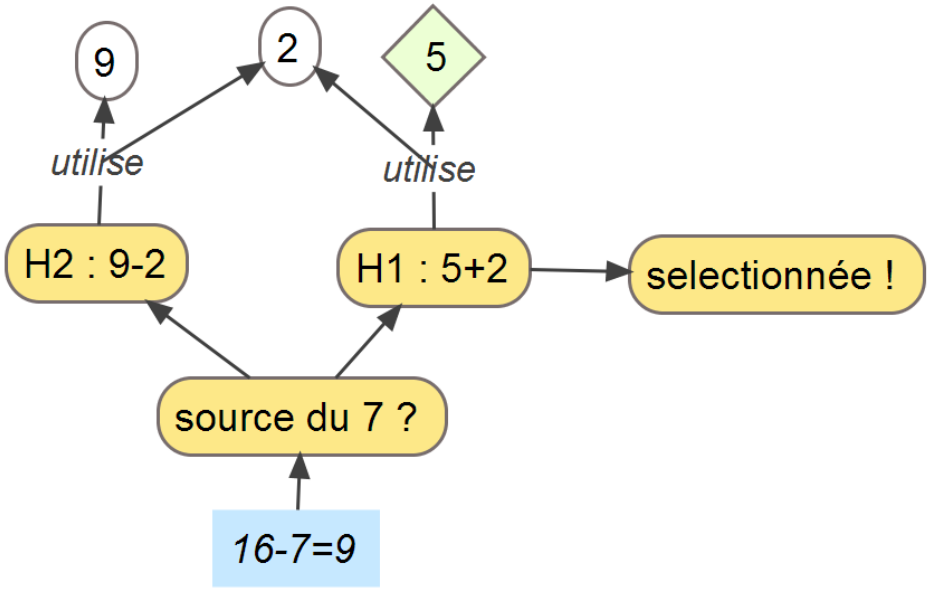



Figure 19. Sélection de la source la plus probable pour un calcul mental. Contrairement à l'hypothèse H2, H1 utilise le dernier nombre calculé par le sujet (5), c'est donc la source qui est sélectionnée.

Finalement, la représentation synthétique de la réponse de l'élève est ((14-9)+2)+9=16

### 7.2.5 Enregistrement des prises de risque

Au cours de la lecture des parties précédentes, la sensibilité des opérations de désambigüisation des formules du programme peut sembler inquiétante. N'est-il pas préférable de suspendre le diagnostic plutôt qu'utiliser des règles invitant le programme à faire des choix dans l'incertain ? Plutôt que de déclarer des formules ininterprétables quand des cas difficiles se présentent, nous avons choisi de formaliser le fait que certaines opérations effectuées par les modules ont tendance à faire baisser le degré de confiance dans le diagnostic. Nous avons décidé de prendre en compte ces cas pour

attribuer un score de « **non-fiabilité** » du programme. Voici la liste complète des cas considérés comme problématiques et qui incrémentent le compteur de non-fiabilité :

Du point de vue des formules :

- Une formule est considérée comme ininterprétable.
- Trop de calculs mentaux sont possibles pour expliquer une formule à désambigüiser : cas que nous avons appelé « formule saturée » précédemment.

Du point de vue des nombres :

- Plusieurs calculs mentaux sont possibles pour expliquer un même nombre. Le compteur est augmenté une fois de plus si le programme n'est pas arrivé à prendre de décision.
- Lorsque plusieurs sources sont possibles pour expliquer un nombre (issu d'un calcul, de l'énoncé ou d'un résultat d'une reprise fautive du nombre après le signe égal)
- Un nombre dans la réponse est d'origine inconnue, ni calculée, ni issue de l'énoncé.

Si ce compteur prédit correctement la fiabilité du diagnostic, il peut avoir un intérêt pratique important dans une logique de collaboration personne-machine. Lorsqu'un désaccord se produit entre le diagnostic manuel et le diagnostic automatique, selon la valeur du compteur, l'humain peut être invité à corriger le diagnostic automatique ou corriger son propre diagnostic manuel (détection d'erreurs).

Le professeur des écoles peut aussi bénéficier de cette mesure. Lorsque le diagnostic de ses élèves lui est fourni, il peut être associé à un pourcentage indiquant la fiabilité du diagnostic.

## 7.3 Validation du diagnostic comportemental dans DIANE

Le module de diagnostic a été testé sur les données que nous avons décrites à la suite de la problématique. Un diagnostic comportemental manuel a été réalisé par l'expérimentateur pour permettre l'analyse des données. Rien ne garantit que ce diagnostic soit meilleur que le diagnostic cognitif automatique réalisé par notre programme, rien ne garantit non plus que toutes les réponses ont un diagnostic comportemental objectif dans la mesure où un travail d'interprétation est réalisé. Nous reviendrons sur ces points dans le cadre d'une expérimentation présentée plus loin.

Toutefois, cette expérimentation contient peu d'observations et ne peut pas intervenir dans le type d'analyse que nous présentons ci-après. Pour les analyses que nous présentons ci-dessous, nous prenons le codeur humain comme « gold standard ». C'est aussi sur ces données qu'ont eu lieu les premiers tests du programme de diagnostic. Lors de la construction du programme du diagnostic, nous analysions seulement les 50 premiers désaccords entre l'ordinateur et l'humain[25].

### 7.3.1 Comparaison du codage automatique et du codage humain

Pour comparer ces deux types de codage, nous nous sommes concentrés sur l'unique colonne « calcul final » dans la grille de codage fournie dans les données. Elle indique, soit le calcul qui a mené au nombre donné en réponse, soit, par défaut, le dernier calcul effectué. Ces formules sont écrites de manière symbolique à partir des nombres de l'énoncé (T1, P1, d). Lorsque plusieurs étapes sont réalisées, elle est écrite de manière composée (e.g la formule (T1-d)+(P1-d) résume trois calculs T1-d=x, P1-d=y, x+y=z). C'est donc la colonne la plus importante du codage, car elle contient la valeur calculée ainsi que les étapes par lesquelles l'élève semble être passé. C'est aussi la plus sensible puisque si une étape n'est pas bien diagnostiquée, la formule finale ne peut pas être juste. Il est donc pertinent de ne comparer les diagnostics que sur cette formule finale. Tout comme le diagnostic automatique, lorsque l'expérimentateur n'a pu déterminer une interprétation satisfaisante, il note que la réponse est « *ininterprétable* ».

L'établissement de cette comparaison s'est heurté au problème des conventions dans le codage des réponses. Elle peut être laissée vide si un calcul mental semble être à l'origine du résultat. L'humain et le programme peuvent donner des formules équivalentes, mais pas identiques du point de vue de la comparaison brute de la chaîne de caractères.

[25] Au total 200 désaccords spécifiques ont été étudiés dans cette phase de conception. Leurs études ont permis d'identifier les premières sources de désaccords qui pouvaient être gérés par un système de décision de l'ordinateur. Cette méthode occasionne peu de risque de surajustement, car l'observation des désaccords est générale et a guidé la production de modules généraux.

Voici quelques exemples :

- T1+d versus d+T1
- (T1-P1)+(T1-d) versus (T1-d)+(T1-P1) ou T1-d+(T1-P1)

Comparer des chaînes de caractères ne suffit donc pas pour établir l'équivalence de deux formules mathématiques. Pour établir l'égalité (ou la non-égalité) des formules, nous avons utilisé une approche mixte basée sur :

- Le comptage des symboles, car pour que deux formules soient équivalentes, elles doivent avoir le même nombre de T1,P1 et d.
- L'évaluation numérique des expressions lorsque T1 est remplacé par 100, P1 par 10, et d par 1. Si ces évaluations diffèrent, c'est que l'expression est différente.
- L'utilisation des parenthèses. Si les deux formules ont le même nombre de parenthèses, alors une évaluation numérique est effectuée à l'intérieur de chaque parenthèse (cf. point 2). Les valeurs sont comparées pour s'assurer que la forme est la même. Cette technique permet d'éviter les faux positifs (par exemple : (P1+T1)+d doit être considérée différente de P1+(T1+d), car l'étape intermédiaire n'est pas la même).

Sur les 4532 réponses constitutives de la base de données fournie, nous avons obtenu 4000 accords (exactement) et 532 désaccords soit environ **88 % d'accord** entre le codeur humain et le codage automatique. Le codage automatique est donc plutôt satisfaisant. Il est possible de l'utiliser dans le retour au professeur en indiquant qu'il est en accord avec l'humain dans cette même proportion.

Cependant, est-il suffisamment fiable pour remplacer le codeur humain dans le cadre d'une recherche expérimentale en psychologie ? La question reste ouverte. Si l'échantillon est suffisamment large, une part de bruit dans le codage peut être acceptable, mais s'il est restreint, un chercheur pourrait ne pas vouloir se contenter de 88 % de fiabilité. Par ailleurs, il est possible que les 12 % de désaccords puissent être le fait d'erreurs humaines. Ainsi le programme de diagnostic pourrait être utilisé comme système de confirmation du codage humain, en soulignant les cas qui mériteraient une deuxième analyse, éventuellement par un deuxième codeur, pour améliorer la confiance dans le codage.

### 7.3.2 Analyses de la répartition des désaccords avec le codeur humain

Un premier niveau d'analyse disponible est l'analyse des réponses que le diagnostic automatique déclare ininterprétable. La question est alors : « est-ce que le codeur humain déclare lui aussi que ces réponses sont ininterprétables ou bien propose-t-il une interprétation de cette réponse ? » Selon les données reportées dans le Tableau 4, un quart des désaccords (532/122) provient de l'échec d'interprétation du programme qui déclare la réponse comme ininterprétable. Réciproquement, lorsque le diagnostic automatique indique qu'une réponse est ininterprétable la probabilité que le codeur humain soit en désaccord avec ce diagnostic est relativement élevée et dépasse les 50 % (122/214).

Tableau 4. Répartition des observations selon le statut interprétable ou ininterprétable donné par la machine et l'accord ou le désaccord entre l'humain et la machine.

| Statut selon la machine | Désaccord | Accord | Sous-total |
|---|---|---|---|
| ininterpretable | 122 | 92 | 214 |
| interprétable | 410 | 3908 | 4318 |
| Sous-total | 532 | 4000 | 4532 |

D'un point de vue pratique, si l'effort humain est limité, alors il peut ne réaliser le diagnostic manuel que sur les « ininterprétables » du modèle. Ces ininterprétables représentent seulement environ 20 % des données (214/4532) et détiennent un haut taux de désaccord (122/214).

Nous rappelons que nous avons mis en place un compteur pour mesurer la non-fiabilité du diagnostic. Le Tableau 5 représente sa répartition croisée avec l'accord humain-ordinateur :

Tableau 5. Répartition des accords/désaccords selon la valeur du compteur de non-fiabilité.

| valeurs du compteur de non-fiabilité | accord Homme-Machine | désaccord Homme-Machine |
|---|---|---|
| 0 | 3537 | 257 |
| 1 | 383 | 126 |
| 2 | 46 | 64 |
| 3 | 20 | 50 |
| >3 | 14 | 35 |

Lorsque le compteur est à 0, les cas de désaccord sont rares (évalués à 7 %). Lorsqu'il augmente, le pourcentage croît brutalement : 24 %, 58 %, 71 % pour les valeurs

respectives du compteur de 0,1,2,3. **Les opérations risquées dans le diagnostic sont de ce fait bien identifiées et la mise en place de cette mesure est donc pertinente**. En cumulé, lorsque le compteur est supérieur à 0, nous obtenons 37 % de désaccord avec le codeur humain.

Nous avons donc deux indicateurs de la non-fiabilité du diagnostic : le compteur de non-fiabilité et le caractère interprétable (ou non) de la réponse selon le programme. Les deux prédicteurs de non-fiabilité sont-ils cumulables, ou peut-on se passer de l'un des deux ?

D'un point de vue descriptif, l'examen du Tableau 6 semble répondre en faveur de la compatibilité des deux informations.

Tableau 6. Répartition des accords et désaccords selon la valeur du compteur de non-fiabilité et selon le statut interprétable ou ininterprétable donné par la machine

| statut selon la machine | valeurs du compteur de non-fiabilité | accord Homme-Machine | désaccord Homme-Machine |
|---|---|---|---|
| ininterpretable | 0 | 30 | 30 |
| ininterpretable | 1 | 22 | 28 |
| ininterpretable | 2 | 19 | 20 |
| ininterpretable | 3 | 11 | 30 |
| ininterpretable | >3 | 10 | 14 |
| interpretable | 0 | 3507 | 227 |
| interpretable | 1 | 361 | 98 |
| interpretable | 2 | 27 | 44 |
| interpretable | 3 | 9 | 20 |
| interpretable | >3 | 4 | 21 |

Deux remarques peuvent être faites : lorsqu'une formule est déclarée ininterprétable, il semble que le compteur perd de son intérêt et que le taux d'accord stagne autour de 60 %, quelles que soient ses valeurs ; à l'inverse, lorsqu'une formule est déclarée interprétable, le compteur joue un rôle fondamental pour prédire les désaccords. Il y a donc une compatibilité forte entre les deux indicateurs.

### 7.3.3 Analyse des performances du diagnostic module par module

La construction du programme de diagnostic fait appel à de nombreux mécanismes. Le taux d'accord entre le codeur humain et les productions du programme suggère une certaine efficacité d'un point de vue global, mais ne permet pas de savoir si les

mécanismes sont pertinents un à un. Cette partie vise à répondre à cette question d'un point de vue quantitatif. Nous avons analysé l'évolution du pourcentage d'accord entre le programme et le codeur humain lorsqu'un mécanisme est supprimé ou remplacé par un processus aléatoire lorsqu'il s'agit d'une décision entre plusieurs hypothèses. Nous avons donc rajouté à notre modèle des paramètres globaux visant à altérer le comportement de certains modules. Les mécanismes étudiés sont les suivants :

- Mécanisme de modification des phrases de la réponse de l'apprenant (le module **AG_1**).
- Mécanisme d'inférence de calcul mental (utilisé dans **AS**).
- Mécanisme de suppression de calcul mental dans le cas où la formule est saturée (utilisé dans **AS**)
- Mécanismes de sélection des calculs pour expliquer la provenance d'un nombre dans le cas où des possibilités contradictoires coexistent.

Le Tableau 7 compte l'**augmentation du nombre de désaccords entre l'humain et l'ordinateur** par rapport à la valeur de base du modèle complet. Avec plus de 40 % d'augmentation du nombre de désaccords avec l'humain par rapport au modèle complet, il est clair qu'on peut difficilement se passer des trois premiers mécanismes.

Tableau 7. Augmentation des désaccords lorsqu'un mécanisme est supprimé ou lorsque son processus de décision est remplacé par un choix aléatoire.

| mécanisme altéré | suppression | aléatoire |
|---|---|---|
| Modification des phrases | + 43,75% | - |
| Mécanisme d'inférence de calcul mental | + 72,25% | - |
| Mécanisme de selection de calcul mental lorsque la formule est saturée (tous les nombres peuvent être expliqués par calculs mentaux) | + 51,83% | + 28,17% |
| Mécanisme de selection de calculs mentaux lorsque plusieurs calculs expliquent le même nombre | +7,58% | + 0,83% |
| Mécanisme de selection de calculs explicites lorsque plusieurs calculs expliquent le même nombre | + 15% | + 2,75% |

Le constat est plus mitigé pour les mécanismes de sélections portant sur la sélection de la provenance d'un nombre. Ces deux mécanismes ont un mécanisme de décision commun[26]. Les pourcentages relativement importants dans la colonne « *suppression* » nous indiquent l'importance de ne pas suspendre l'interprétation du calcul lorsque ces cas se présentent, mais la colonne aléatoire nous indique que nos principes de sélection ne valent guère mieux que le hasard.

Pourquoi ces deux modules donnent-ils de mauvais résultats ? La première cause possible analysée est la pertinence de la liste ordonnée classant les différentes sources possibles d'un nombre allant du plus probable au moins probable. Une autre possibilité est que certains éléments dans la réponse du sujet ne sont pas pris en compte dans le diagnostic automatique soient pris en compte par le codeur humain.

## 7.3.4 Codage automatique pour corriger le codage humain

### 7.3.4.1 Le diagnostic automatique peut-il efficacement corriger l'humain ?

Précédemment, nous avons temporairement considéré le codage humain comme un « gold standard ». Mais rien n'exclue que l'humain puisse faire des erreurs, et dans ce cas, le codage automatique peut-il s'avérer un bon moyen pour repérer les erreurs humaines ?

[26] Plus précisément le mécanisme de sélection de calculs mentaux utilise et somme pour chacun de ses opérandes le système de score établi par le mécanisme de sélection de calculs explicites

Lorsque nous étudions les désaccords, nous pouvons noter qu'ils présentent une certaine diversité. Certains désaccords donnent raison à l'humain et pointent des limites semblant infranchissables du diagnostic automatique. Le premier exemple dans le Tableau 8 fait partie de ces cas : ici l'ordinateur déclare la réponse comme ininterprétable. En effet, le tout premier calcul n'est pas compréhensible. Face aux données de l'énoncé, le codeur humain, lui, a compris que 98 n'était rien d'autre que 9 et 8. 8 (ou 9) étant une donnée de l'énoncé, tout fait sens s'il est supposé que l'élève a commis une faute de recopiage. Ce niveau de déduction humaine pourrait difficilement être traduit par des règles. Il serait possible de faire une règle simple comme « *si un nombre à deux chiffres apparait dans une formule incompréhensible, alors essayer de séparer les nombres pour désambigüiser la formule* ». Cela fonctionnerait sur cette réponse spécifique, mais il nous est difficile, par ce nouvel ajustement, d'estimer les faux positifs qu'il causerait. Nous préférons en effet que le diagnostic indique un faux négatif (« ininterprétable ») plutôt qu'un faux positif.

Tableau 8. Quelques désaccords entre l'homme et la machine

| Problème | Réponse | codage machine | codage humain |
|---|---|---|---|
| Quand Médor monte sur la balance chez le vétérinaire, la balance indique 9 kilos. Quand Médor et Rex montent ensemble sur la balance chez le vétérinaire, la balance indique 15 kilos. Fido pèse 4 kilos de moins que Médor. Fido et Rex montent ensemble sur la balance chez le vétérinaire. Combien Fido et Rex pèsent-ils ensemble ? | 15-98=6 9-4=5 6+5=11. Fido et Rex pèsent 11 kilos ensemble | ininterprétable | (T1-P1)+(P1-d) |
| | 6+3=9. Fido et Rex pèse 9 kilo ensemble | ininterprétable | T1-P1)+(P1-(T1-P1)) |
| Dans la classe de CM2, il y a 9 élèves. Si on réunit les CM2 et les CM1, cela fait un groupe de 17 élèves. Dans la classe de CE2, il y a 3 élèves de moins qu'en CM2. On fait un groupe réunissant les CE2 et les CM1. Combien y a-t-il d'élèves dans ce groupe ? | 8+5=13. Le groupe de CM1 et CE2 font 13 élève | ininterprétable | (T1-P1)+(P1-d) |
| En janvier, 6 enfants se sont inscrits à la chorale. Après janvier, il y a 13 enfants à la chorale. Avant janvier, il y avait autant d'enfants inscrits au football qu'à la chorale. En janvier, il y a eu de nouvelles inscriptions au football. Après janvier, il y a 2 enfants de moins au football qu'à la chorale. Combien d'enfants se sont inscrits au football en janvier ? | Tt3_v2 6-2=4 4+6=10. Il a 10 enfant inscri au football | (P1-d)+P1 | (T1-d)+P1 |

Certains désaccords sont parfois difficiles à départager, des arguments existent dans les deux camps, c'est le cas pour le deuxième exercice de la table. Ici, il y a matière à débat, l'humain a peut-être de bonnes raisons de penser qu'un enchaînement de deux calculs mentaux est à l'origine du 3 dans la réponse « 6+3=9 ». Mais peut-on en être sûr ? Le

programme, lui, ne peut désambigüiser une formule que par un calcul mental pour éviter de prendre des risques. La troisième formule présentée ici est un cas voisin. Ici l'humain a considéré qu'il y a eu calcul mental avec erreur de calcul, ce que ne fait pas notre programme pour éviter la multiplication des sources possibles pour un nombre. Il est difficile d'avoir un avis tranché sur ces formules. D'un côté, l'humain peut avoir de bonnes raisons auxquelles l'algorithme n'a pas accès pour faire ses inférences (par exemple si ce sont des calculs qui sont habituellement faits par l'enfant, ou si l'enfant est bon sur les autres problèmes...etc.). Mais d'un autre côté, l'humain peut-il efficacement contrôler la présence d'explications alternatives lorsqu'il fait ses inférences ? Le nombre de chiffres accessibles par double calcul mental est sans doute possible, assez large avec une certaine quantité de possibilités menant aux mêmes nombres.

Des formules, par contre, semblent remettre en cause l'humain. C'est du moins notre interprétation qui reste à être confirmée. La dernière réponse dans le tableau est une de ces erreurs. Au vu de ce désaccord, il semble que l'humain ait simplement effectué une erreur d'inattention et remplacé P1 par T1.

Pour quantifier le niveau de discussion ou le niveau de remise en cause que la machine peut produire dans le codage humain, nous avons construit une expérimentation. Outre avoir une estimation sur la quantité relative d'erreurs humaines dans les cas de désaccord, nous souhaitons utiliser notre compteur de confiance pour affiner cette estimation. Sur la base de l'évolution du niveau de désaccord, nous faisons l'hypothèse que la proportion de fautes humaines détectée est plus forte pour la catégorie dans laquelle le compteur vaut 0 que pour les autres catégories (>0). Nous faisons aussi l'hypothèse que les jugements sont plus incertains pour l'humain dans ces catégories supérieures, l'idée étant de montrer que l'humain est aussi sensible à ce que l'algorithme relève comme problématique dans la réponse de l'apprenant. L'incertitude dans le jugement sera mesurée de deux manières :

- La confiance dans le jugement sera modélisée par un indicateur prévu à cet effet.
- L'accord inter juge sur ces jugements.

### 7.3.4.2 Matériel

#### *7.3.4.2.1 Données récoltées*

Nous avons sélectionné au hasard 100 désaccords et avons produit une grille dont nous donnons les cinq premières lignes. La passation est informatique, les feuilles sont ouvertes avec un tableur. Elle contient le code du problème, les nombres de l'énoncé la réponse de l'enfant.

| idPbm | protocole | T1\|P1\|d | formule finale - codeur 1 | | formule finale - codeur 2 | | confiance | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ct4_v2 | 7-3=4 7+4=11. 11€ | 15\|7\|3 | P1+(P1-d) | | ininterp | | 1 | 2 | 3 | 4 |
| Ct2_v1 | 6-2=4 kilos 15-6=09 Fido pèse 4 kilos et Rex pèse 9 kilos | 15\|6\|2 | T1-P1 | | P1-d | | 1 | 2 | 3 | 4 |
| Ct1_v2 | 16-7-4=7 16-7-4=7. Il auron 7 billes | 16\|7\|4 | T1-P1-d | | (T1-((T1-P1)-d))-d | | 1 | 2 | 3 | 4 |
| Ct1_v1 | (5-3 : effacé) Jaque : 2 billes Paul : 7 billes. 7+2=9 ils ont 9 billes emsemble | 12\|5\|3 | (T1-P1)+(P1-d) | | (P1+(P1-d))+(P1-d) | | 1 | 2 | 3 | 4 |
| Ct2_v2 | 9-4=5 15-9=6 9+6=15 5+6=11. Ils pèsent 11 kilos. (9-4=5 ; 15-9=6). | 15\|9\|4 | (P1+(T1-P1))-P1 | | (T1-P1)+(P1-d) | | 1 | 2 | 3 | 4 |

Le codage de l'humain et le codage machine sont randomisés, les participants ne savent donc pas dans quelle colonne est située le codage ordinateur.

#### *7.3.4.2.2 Consignes*

Il est demandé au codeur d'entourer ce qu'il pense être le bon codage. À gauche, le niveau de confiance dans son choix est demandé. Les cases possibles sont : (1) confiance faible, il y a matière à débat ; (2) le choix peut éventuellement être discutable, mais le diagnostic choisi semble meilleur ; (3) assez sûr de son choix ; et (4) complètement sûr : l'un des diagnostics a commis une erreur.

Pour faciliter leurs prises de décisions, les nombres des énoncés sont reportés. Les énoncés des problèmes sont aussi fournis sur une feuille à part, ainsi les codeurs peuvent les consulter par le biais du code du problème fourni en première partie.

Pour éviter tout malentendu, quelques conventions de codage sont rappelées dans les consignes (en annexe), et il est expliqué que la place du codeur humain et du codeur automatique est randomisée et que les réponses qui se suivent ne sont pas du même élève (sauf par extrême chance)

#### *7.3.4.2.3 Participants*

Les participants sont 3 codeurs experts. Le premier est l'auteur du manuscrit, le deuxième est le codeur humain de la base de données présentée à la suite de la problématique et la troisième experte a déjà réalisé ce type de diagnostic sur d'autres corpus de données.

*7.3.4.2.4 Codage*

Nous construisons une variable nommée « votes en faveur de l'humain » qui compte combien de votes l'humain a obtenu en faveur de sa formule. Elle varie donc entre 0 et 3. Lorsque la valeur est de 0 ou 1, il est considéré que le codage automatique obtient la majorité. Lorsque sa valeur est de 2 ou 3, c'est le codage humain qui obtient la majorité.

7.3.4.3 Résultats de l'expérimentation

En annexe, des liens internet peuvent mener au détail de l'analyse statistique de cette expérimentation (ainsi que les analyses précédentes sur le diagnostic comportemental et son code source).

Les résultats principaux sont reportés dans le Tableau 9.

Tableau 9. Répartition des votes en faveur de l'humain en fonction de la valeur du compteur de non-fiabilité.

| valeurs du compteur de non-fiabilité | votes en faveur la formule de l'humain | | | |
|---|---|---|---|---|
| | 0 | 1 | 2 | 3 |
| 0 | 14 | 12 | 7 | 14 |
| >0 | 3 | 8 | 13 | 29 |

Lorsque l'humain obtient 2 ou 3 votes en sa faveur il est considéré que le jugement du codeur humain l'emporte, dans le cas contraire, c'est le diagnostic automatique qui l'emporte. Le nombre de réponses dont le désaccord personne-machine est tranché en faveur de la machine est de 37 sur 100. En utilisant la méthode asymptotique de Wald[27] permettant la construction d'intervalle de confiance pour les proportions, nous obtenons un intervalle de confiance entre 28 % et 47 % de désaccords tranchés en faveur du diagnostic machine.

[27] D'autres méthodes existent (Clopper-Pearson, Wilson, Agresti-Coull, Jeffreys). L'arrondi au pourcentage près donne ici le même résultat pour les trois méthodes.

Nous pouvons, par la lecture du Tableau 9, attester de la valeur du compteur de confiance pour augmenter les doutes sur le diagnostic de l'humain ou sur son propre diagnostic. En effet, lorsque le compteur est à 0, il obtient 26 délibérations en sa faveur contre 21. Au contraire, lorsqu'il atteint des valeurs supérieures, il n'obtient que 11 délibérations contre 42 en faveur de l'humain.

Un test du Chi2 nous permet de rejeter l'hypothèse d'indépendance : Chi2= 14.643, df = 3, p-value < 0.01.

Ces résultats nous permettent de réestimer à la hausse les capacités du diagnostic automatique. Nous comptions 257 désaccords contre 3537 accords lorsque le compteur est à 0. En utilisant la borne pessimiste de l'intervalle de confiance calculé plus tôt, nous pouvons considérer que ce nombre de désaccords représente au moins 170 « erreurs »[28] de l'humain ce qui permet d'assurer que plus de 95 % du diagnostic automatique sont fiables lorsque ce compteur est à 0.

## 7.3.5 Discussion

Nous réétudions dans cette partie l'hypothèse de la possibilité de construction d'un diagnostic comportemental générique.

### 7.3.5.1 Généralisation des résultats

Les données qui ont servi d'évaluation du diagnostic automatique sont issues d'un livret papier et sont réécrites à la main dans un tableur. Les conditions sont donc très différentes d'une récupération des données directement au sein de DIANE. Dans l'environnement, l'écriture d'une réponse est guidée par des fonctionnalités (pavé numérique, clics sur les mots) et est plus contrainte que l'écriture manuelle d'une réponse. Les données devraient alors être plus faciles à analyser lorsqu'elles sont produites dans DIANE. Lorsque le codeur humain a enregistré les protocoles en les

[28] Le terme « erreur » ici est un raccourci de langage, et il convient de rappeler que pour beaucoup de diagnostics, l'incertitude sur leur validité reste présente. Nous considérons qu'un diagnostic est erroné si les votes des codeurs humains ne sont pas en sa faveur. Or, les codeurs humains sont rarement en accord parfait dans leurs votes, surtout pour les cas de désaccord lorsque le compte de fiabilité est à 0.

retapant, le problème de l'encodage des dessins et des schémas de manière verbale s'est produit et donne des informations encodées de manière particulière. Voici un exemple :

> *dessin : À 7 billes+ P 9 billes=16 et dessin des 16 billes. P 9billes+ J 3billes=12 puis dessin des 12billes. 3 et dessin des 3 billes*

Notre module ne peut pas comprendre ce type de réponse et est donc mis en échec. 94 des réponses de notre ensemble de données mettent en jeu des schémas et des dessins, ce qui n'est pas négligeable. Nous n'avons pas exclu ces données, car il est important de compter ces cas que le module de diagnostic ne peut pas traiter. Toutefois, DIANE ne permettant pas de schématiser un problème, nous pouvons estimer à la hausse les résultats du diagnostic. Il est aussi important de rappeler que les problèmes traités ici sont relativement complexes. Il est possible que les résultats soient meilleurs sur des problèmes simples dans la mesure où moins de calculs mentaux doivent être inférés, ce qui représente la difficulté principale de notre programme[29].

7.3.5.2 Poids supérieur accordé aux bonnes réponses

Il n'y a pas de principe de charité, qui en cas de doute, sélectionne le calcul correct. La question est ouverte s'il fallait ou non prendre cela en compte dans l'arbitrage que nous avons détaillé précédemment. Cette question a de multiples facettes. Tout d'abord cette décision impliquerait une perte de généricité, bien que mineure, dans le programme de diagnostic puisqu'on doit renseigner la bonne réponse du problème et les différentes stratégies pouvant y mener. Le deuxième problème est la quantification de cet avantage qu'on laisse aux bonnes réponses, outre gérer des cas simples comme les cas d'égalité, la rendre compatible avec les autres politiques de décision n'est pas forcément simple.

Des arguments existent pour le camp opposé : Le rapport de coût entre « attribuer une erreur à tort » et « oublier une erreur » n'est pas peut-être pas le même dans une situation pédagogique (on appliquerait alors plus volontiers le principe de charité) que

[29] Les précédentes versions du diagnostic (Hakem, Chaillet, & Sander, 2011) vont dans le sens de notre supposition en reportant de meilleurs résultats pour les problèmes simples que pour les problèmes complexes.

dans une recherche en modélisation cognitive. Dans ce dernier, la question de recherche peut aussi faire varier les décisions. D'un point de vue pragmatique, nous pensons qu'intégrer une connaissance sur les différentes stratégies de résolution pourrait rapprocher significativement l'accord humain-ordinateur. Par exemple un cas de désaccord extrêmement fréquent dans nos données est :

- L'humain détecte que le problème est résolu par la stratégie par étapes (T1-P1)+(P1-d)
- L'ordinateur détecte qu'il a été résolu par différence (T1-d).

Il est clair que l'ordinateur est incapable de comprendre la signification des nombres dans la réponse, car il n'infère des calculs que lorsqu'il détecte des formules. L'humain lui, prend les indices qui lui permettent de considérer qu'il a effectivement établi une stratégie par étape, par exemple la présence de nombre isolé dans la réponse. Si l'*a priori* qu'un élève passe par ces trois calculs (T1-P1), (P1-d) finalisés par (T1-P1)-(P1-d) n'était pas si fort, alors nous pourrions douter que l'humain fasse cette inférence. Cette inférence est tout à fait valide d'un point de vue bayésien, cadre dans lequel le support des données pour estimer la probabilité d'un évènement est multiplié par sa probabilité d'occurrence *a priori*. Nous notons que cette réflexion peut être mise en regard du critère de DPF que nous n'avons pas utilisé dans nos prises de décision : le critère « Minimal Error » que nous avons présenté dans la partie théorique. Il serait donc souhaitable, en perspective, d'implémenter ce critère et d'observer s'il permet d'obtenir un accord meilleur avec le codeur humain.

Dans la poursuite de cette réflexion, les résultats modestes de notre liste ordonnée de décisions pourraient être améliorés en accordant un poids plus fort aux hypothèses qui recouvrent des cas fréquents. Il serait par exemple possible d'associer les nombres isolés (c'est à dire non intégrés à une formule déjà détectée) à un calcul si ce dernier correspond à une stratégie de résolution fréquente.

#### 7.3.5.3 Contributions méthodologiques

Dans une logique de recherche en EIAH fortement couplée à des problématiques de recherches en psychologie expérimentale, la validité des diagnostics comportementaux établis est fondamentale. Plus que la conception d'un système de diagnostic, nous avons construit une méthodologie pour construire et valider un système de diagnostic comportemental. Si les principes permettant de développer un tel système sont déjà bien

connus (Langley et al., 1990), l'étude scientifique du système produit est généralement délaissée[30]. Notre méthodologie consistait successivement à :

- Rendre modulable notre système pour pouvoir étudier un à un la pertinence des modules incarnant des principes de décisions.
- Faire remonter les prises de risque par un compteur de fiabilité.
- Etablir la pertinence du diagnostic en comptant les désaccords avec le codeur humain (en croisant les analyses avec les deux premiers points précédents).
- Investiguer plus en détail la nature des désaccords et la pertinence du compteur de fiabilité par le biais d'une expérimentation et demander à plusieurs juges de voter.

## 7.4 Discussion sur la généricité de DIANE

### 7.4.1 Diagnostic des réponses

Nous pouvons noter que cette modélisation des réponses ne nécessite pas une connaissance des problèmes auquel répondent les élèves si ce n'est les nombres qu'ils comportent. Les informations prises en compte se réduisent aux nombres de l'énoncé, à la réponse de l'élève.

Ainsi, le diagnostic effectué est tout à fait générique et en accord avec l'évolution de DIANE dans ses nouveaux modes de conception de problème. Son extension aux problèmes de multiplication est possible à condition que les modules de détections de formules soient mis à jour et que les modules de traitement de formules soient aussi prêts à identifier les différents types ou formes de multiplication/division pour que l'approche fasse sens. Elle serait souhaitable, car nous étions en présence d'un certain nombre de réponses comportant des multiplications (18) et des divisions (20) qui sont

[30] Une connaissance exhaustive de la littérature nous est impossible, il est donc possible qu'une approche semblable ait déjà été produite sans nous être parvenue. Les précédentes études du diagnostic de DIANE (Hakem, Sander, Labat, & Richard, 2005) pourraient former une exception, mais seul le comptage de désaccord est présent. Les trois autres points de la méthodologie que nous proposons sont absents.

donc ininterprétables par le programme. Mais se pose alors la question de l'arbitrage. Ajouter des multiplications et divisions augmente les possibilités de désambigüisation des formules, car plus de calculs mentaux sont accessibles. Nous suggérons donc, lorsqu'un exercice ne contient pas de structures multiplicatives, de ne pas prévoir des calculs mentaux de ce type pour éviter les faux diagnostics.

Du point de vue des performances du diagnostic, les performances dans l'ancienne version de DIANE étaient notées comme très bonnes : pour chaque colonne de l'ancienne grille de codage, un taux d'accord d'environ 95 % avec le codage humain était relevé. Toutefois, comme nous l'avons soulevé plus tôt, ce diagnostic était optimisé pour un certain type de problèmes et ne relevait qu'un certain type d'indicateurs. Il souffrait par ailleurs d'un manque de généricité et de modularité. La perte de précision occasionnée par cette évolution du module de diagnostic nous semble être un prix à payer raisonnable.

### 7.4.2 Dissociation entre modélisation cognitive et modélisation comportementale

Nous nous sommes placés dans la distinction du diagnostic en deux étapes distinctes : la modélisation comportementale et la modélisation épistémique (ou cognitive par abus de langage). Le but était d'attribuer le moins possible de significations et d'intentions au niveau comportemental. Notre travail, consistant à modéliser le comportement sur des réponses ouvertes dans le micro domaine des PAEV suivait cet objectif. Les performances du programme semblent être satisfaisantes. Cependant, des limites importantes peuvent être mises en avant. Dans la partie précédente, nous avons expliqué qu'il n'y avait pas de principe de charité dans notre programme. En revanche, il est possible de remarquer que le principe de charité est d'une certaine manière appliqué quant à la cohérence des réponses. En effet, nous faisons continuellement l'hypothèse que la suite des calculs fait un sens pour désambigüiser les formules (en regardant avant ou après la formule courante). Sans cette part d'interprétation, beaucoup de formules resteraient ininterprétables. Outre les désambigüisations, l'inférence de l'existence de calculs mentaux frôle la contradiction avec l'objectif de modéliser au niveau comportemental seulement. L'intention, les buts et sous-buts que peut avoir l'élève sont donc modélisés plutôt en accord avec la conception de Wenger sur le diagnostic comportemental. Cependant, elle est modélisée de manière abstraite, car elle n'est produite que par l'idée que les calculs se suivent pour obtenir une réponse finale. Nous

ne sommes donc pas au troisième niveau de modélisation « états mentaux » proposé par VanLehn.

# 8 MODELISATION COGNITIVE. RECHERCHER LES SOURCES D'ERREURS

## 8.1 Introduction

La construction d'un système de diagnostic cognitif a deux étapes. La première est le diagnostic comportemental, c'est ce que nous avons conçu dans le chapitre précédent. La deuxième est le diagnostic épistémique. Pour que ce dernier soit réalisé, il est nécessaire d'avoir des théories opérantes mettant en lien la cognition de l'apprenant et ses productions. Lorsque nous souhaitons quitter le domaine de la prédiction de la justesse et de la fausseté des réponses pour prédire les erreurs, des difficultés apparaissent. De nombreux facteurs de difficultés des PAEV ont été mis en évidence et des modèles développementaux ont fait leur preuve. Cependant, la littérature traite peu de l'analyse des erreurs et les modèles exécutables implémentant les phénomènes étudiés sont relativement rares. De ce fait, la capacité de produire des diagnostics cognitifs profonds dans le domaine des PAEV ne peut être obtenue que par des efforts de modélisation conséquents, transversaux aux différentes approches théoriques que connaît la littérature. Il est possible de voir la progression dans le domaine en deux étapes. La première est d'être capable de prédire les erreurs fréquentes par des composantes explicatives valant pour tous les élèves. Une fois que cette étape est franchie, il est possible de s'intéresser aux éventuelles différenciations des sujets sur les composantes mises en évidence. Il s'agit là d'un travail de longue haleine, qui ne peut

être mené que dans un programme de recherche à part entière. Nos travaux ne peuvent être que parcellaires vis-à-vis de cette ambition.

Dans cette partie, nous reformulons notre analyse de la littérature pour construire, tester et comparer deux modèles pouvant générer des erreurs sur des problèmes à plusieurs étapes. Ce chapitre présente donc un travail en psychologie cognitive, placé de manière intermédiaire entre l'étape de diagnostic comportemental présenté dans le chapitre précédent et l'étape de diagnostic cognitif qui fera l'objet du chapitre suivant.

### 8.1.1 Proposition de modélisation : relâchement progressif de contraintes

Reusser et Stebler (1997) proposent un ensemble de 8 règles dont certaines peuvent être pensées sur la base du contrat didactique. D'autres peuvent être mises en relation avec notre précédente analyse de la littérature portant sur la résolution de problèmes par l'emploi de mots-clefs et sur les difficultés linguistiques. Ces règles nous permettent donc de faire la transition de la présentation de ces phénomènes à leurs modélisations.

v. Prends pour acquis que chaque problème présenté par un professeur ou un manuel fait sens.

vi. Ne remets pas en question l'aspect correct et complet du problème.

vii. Prends pour acquis qu'il n'y a qu'une bonne réponse pour chaque problème.

viii. Donne une réponse pour chaque problème.

ix. Utilise tous les nombres du problème pour trouver la solution.

x. Si l'opération fonctionne sans produire de reste, tu es sûrement sur la bonne voie.

xi. Si un problème se révèle insoluble ou trop difficile, passe par une interprétation évidente du problème passant par les informations du texte et les opérations maîtrisées.

xii. Si tu ne comprends pas le problème, utilise les mots-clefs ou les problèmes résolus précédemment pour déterminer une opération mathématique.

Les quatre premières règles impliquent la production d'une réponse sous la forme d'une opération : pas de réponse vide, et encore moins de remise en question du problème.

Les quatre suivantes guident la réponse de l'apprenant (pas de retenue, utiliser tous les nombres, réinterpréter ou travailler sur la base des mots-clefs si obstacle). Elles peuvent

toutes être traduites comme des contraintes sur la production de réponse. Toutes, sauf la première, sont formalisables en termes de contraintes sur le comportement dans un modèle cognitif de façon plus ou moins simple (règles 7 et 8). Jusqu'à présent, ces aspects n'ont jamais été modélisés en même temps dans le cadre d'un modèle cognitif. Les règles 7 et 8 se traduisent comme des relâchements possibles de contraintes. Si le problème est trop difficile, il est attendu : (1) des réinterprétations, (2) une utilisation des mots-clefs, et (3) une utilisation des problèmes anciennement résolus. Nous avons vu, dans nos chapitres précédents, des travaux soutenant les deux premières hypothèses. Dans cette logique, il est aussi possible de suggérer que les contraintes concernant l'emploi des règles précédentes puissent être relâchées, comme effectuer une opération qui n'est pas facilitée (règle 6), utiliser les nombres de manière non stéréotypée (règle 5) ou éventuellement le fait même de produire une réponse (règle 4). Au vu du comportement observé face aux problèmes absurdes vus précédemment, il est pertinent de considérer que les trois premières contraintes ont un poids très fort, et qu'il faudrait employer des problèmes et des consignes très particulières pour favoriser leur relâchement. Au contraire, les études de Brissiaud (1988) et de Inoue (2005) mettent en avant l'idée qu'un relâchement de contraintes de sens (règle 7) est préféré à la verbalisation d'une remise en cause de l'intégrité du problème. Sur le plan de la modélisation, il est donc pertinent d'utiliser le relâchement de contraintes pour établir des prédictions dans ce type de problèmes.

Knoblich, Ohlsson, Haider et Rhenius (1999) montrent ainsi que la difficulté relative de certains problèmes d'allumettes (« math stick problems ») dépend du poids des contraintes qui doivent être relâchées pour mener à bien la tâche. Bien que ces études soient réalisées dans le cadre de problèmes à Insight, des travaux proposent d'utiliser la théorie des auteurs pour expliquer la difficulté relative dans la décomposition des éléments d'un tout dans les problèmes à énoncé verbaux (Thevenot & Oakhill, 2008).

Sur la base de cette analyse, voici les contraintes que nous suggérons d'étudier :

- Donner une réponse respectant le sens des propositions du problème.
    - Niveau de relâchement 1 : Réinterpréter une proposition du problème.
    - Niveau de relâchement 2 : Se baser sur les mots-clefs.
    - Niveau de relâchement 3 : Ne rien répondre.
- Utiliser tous les nombres.

    - Relâchement : Ne pas utiliser tous les nombres

- Ne pas réutiliser de nombre.
    - Relâchement 2 : Réutiliser un nombre

En comparaison avec les règles de Reusser et Stebler de la sous-partie précédente, notons que:

- La règle d'utilisation des nombres est décomposée en deux contraintes possibles : la non-réutilisation de nombre et l'utilisation de tous les nombres.
- Les trois premières règles (1, 2 et 3) ne sont pas représentées. En effet, nous voulons étudier le relâchement de contrainte, or, les trois premières contraintes possibles (prendre pour acquis que le problème fait sens, est correct et n'a qu'une réponse possible) ont certainement un poids si fort que seuls des problèmes très absurdes pourraient provoquer leur relâchement. Au contraire, nous avons choisi des contraintes dont le poids semble suffisamment faible pour pouvoir être relâchées face à une difficulté dans la résolution.

Nous faisons l'hypothèse que ces contraintes sont préférentiellement respectées, mais peuvent être relâchées si besoin. Nous proposons donc un cadre dans lequel les réponses erronées sont liées à la flexibilité de nos représentations, ces dernières étant conceptualisées par des contraintes. Elles sont de nature diverse et peuvent être relâchées successivement. Certaines sont dites « didactiques », car elles concernent l'emploi des nombres de façon très stéréotypée. D'autres portent sur l'aspect sémantique du problème, et représentent le sens original de chaque proposition. Elles peuvent être relâchées de manière locale, ce qui engendre une modification du sens de la proposition, ou de manière globale. Dans ce dernier cas, c'est l'entreprise même de modélisation de la situation qui est abandonnée, au profit d'une stratégie par mot-clef ou d'un abandon (pas de réponse).

Compte tenu de notre analyse, nous choisissons de nous placer dans un cadre dans lequel les réinterprétations sont, au même titre que le calcul d'une quantité du problème, des opérations élémentaires dans la résolution du problème. Elles sont représentées par des relâchements de contraintes d'interprétation pour, formant alors une interprétation alternative. Cette dernière n'est donc pas forcément le résultat d'une mauvaise compréhension initiale, mais fait partie du processus de résolution des PAEV complexes. Elle s'inscrit comme un changement de représentations tel qu'il est décrit

dans les problèmes à Insight. En considérant le changement de représentation au centre de notre approche, nous nous plaçons dans la continuité des conclusions de Richard et Sander (2000) selon lesquelles la représentation du problème est primordiale et doit être d'un intérêt majeur pour les recherches dans le domaine (p. 14).

> *« La dualité d'interprétations possibles que nous avons mise en évidence passe le plus souvent inaperçue, car pour ceux qui connaissent le problème, il n'y a qu'une interprétation possible, et il faut de longues et patientes recherches expérimentales pour mettre en évidence ces interprétations alternatives chez les sujets en difficulté dans la résolution. On dispose actuellement de suffisamment de données pour qu'on puisse affirmer que la difficulté majeure d'un problème ne réside pas dans la recherche de la solution, mais dans l'établissement de l'interprétation adéquate »*

### 8.1.2 Comparaison avec les approches sans changement de représentation

Cette proposition peut être mise en opposition avec des modèles de résolution retrouvés dans de nombreuses publications expliquant les suites de processus cognitifs intervenant dans la résolution de problèmes arithmétiques. La plupart des modèles conçoivent le processus de résolution en **plusieurs étapes se succédant** (lecture, construction d'une représentation de la situation, construction d'une solution). Nous présentons en Figure 20 quelques-uns de ces modèles. La linéarité de la succession d'étapes peut-être débattue. S'il existe parfois une **boucle retour** dans les modèles graphiques présentés, elle ne concerne généralement que la sélection d'information, ou la vérification que la solution n'est pas incohérente avec les informations données. Toutefois, le modèle graphique de (Hegarty et al., 1992) en Figure 20, prévoit l'apparition d'un traitement du problème par les mots-clefs. Cependant, dans le contexte de l'article, il s'agit plutôt d'une stratégie de résolution qui dépend de l'élève plutôt que d'un changement ponctuel de représentation.

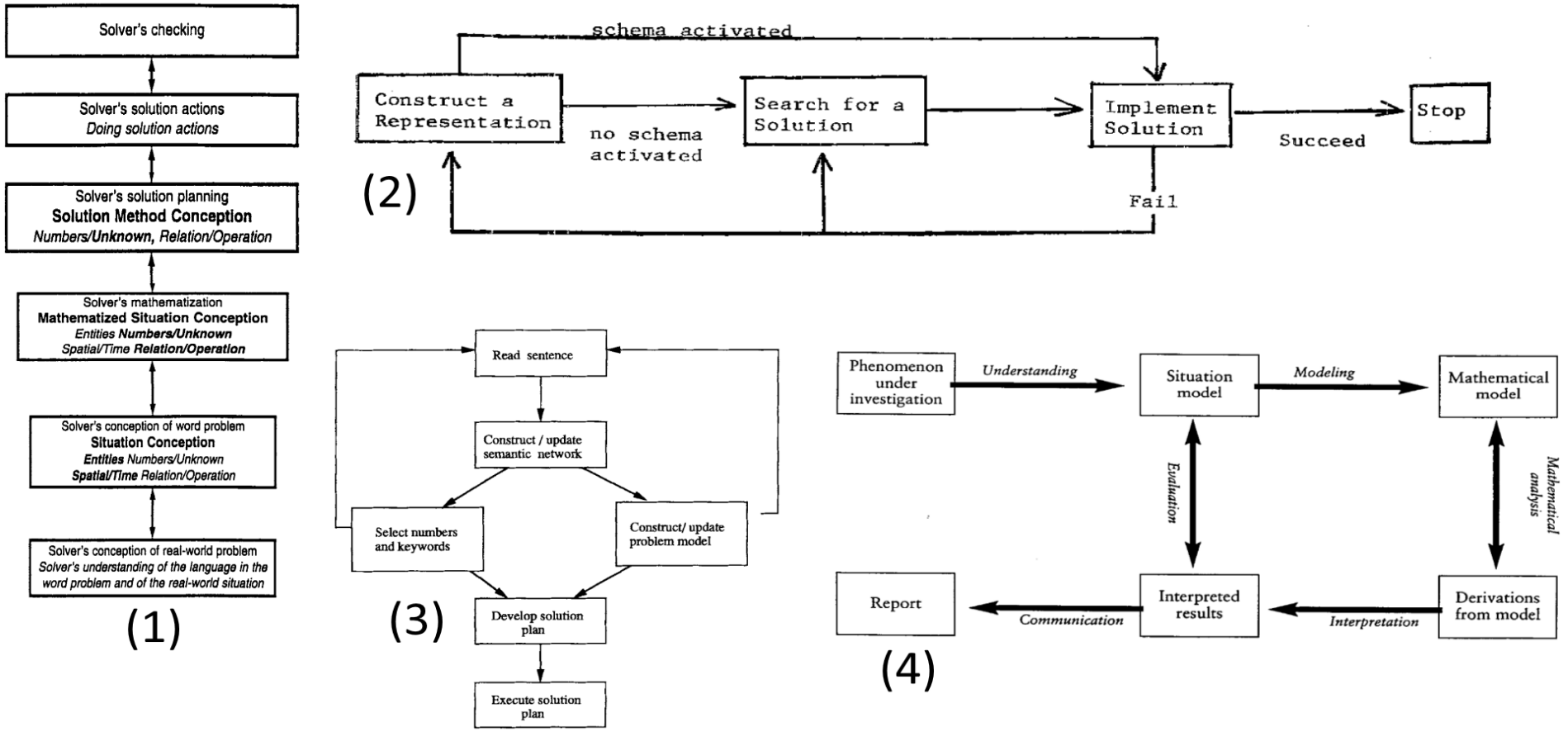



Figure 20 Compilation de représentations graphiques du processus de résolution de problème. 1 : Fuson, Hudson, & Pillar, 1997 (dans Reed, 1999, p. 9); 2 : Gick, 1986 ; 3 : (Hegarty et al., 1995) ; 4 : De Corte, Verschaffel, & Greer, 2000

Dans notre modèle, la boucle retour dont nous voulons montrer l'importance est synonyme de relâchement de contraintes. Cette représentation permet ainsi de mettre au centre les aspects didactiques et les phénomènes de réinterprétation. Nous estimons que (1) ce second travail est avant tout sémantique, les élèves cherchent à donner un nouveau sens aux quantités décrites dans le problème ; (2) il est progressif, les contraintes sont relâchées progressivement.

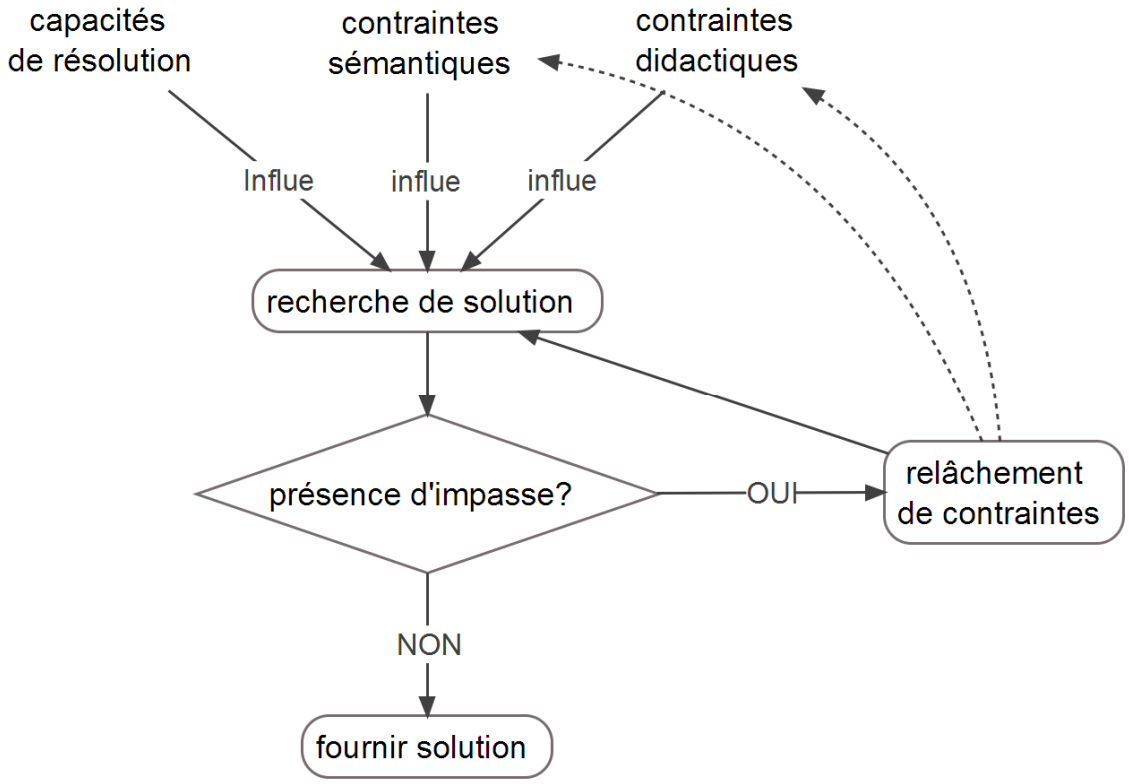



Figure 21. Résolution de PAEV par relâchement de contraintes.

Le schéma en Figure 21 donne une idée générale d'un modèle qui serait centré sur le relâchement de contrainte. Les contraintes sémantiques représentent les propositions contenues dans le texte du problème et leurs relâchements consistent en des réinterprétations. Ce schéma permet de décrire le parti pris théorique qui a guidé la construction de notre modèle. Il est la projection dans le cadre des PAEV de la théorie de résolution problème à l'œuvre dans les différentes implémentations du modèle des contraintes (Richard et al., 2009, 1993). Toutefois, **il n'est possible d'implémenter et de tester qu'une projection de cette théorie générale de la résolution de PAEV**. Cette différence entre « modèle » et « théorie » peut-être ici prise comme analogue à la

différence entre la « Repair Theory » et le programme « Sierra » pour lesquels il est souligné que la théorie est plus large que le modèle, mais seul ce dernier peut être testé (VanLehn, 1990, p. 167).

## 8.2 Construction d'un modèle

Pour appuyer cette idée de relâchement de contraintes dans le processus de résolution, nous avons conçu un programme générant des solutions à des problèmes arithmétiques en se laissant la possibilité de réinterpréter certaines propositions dans le processus. Le but est de constituer une base d'erreurs possibles générée par ce processus et d'étudier son recouvrement avec une base de réponses réelles.

### 8.2.1 Bases du modèle

#### 8.2.1.1 Choix des problèmes

Dans le paragraphe précédent, le processus de résolution de problème a été traduit en termes de contraintes, nous supposons que le sens des propositions est en compétition avec des contraintes d'un autre type comme l'utilisation de tous les nombres. Ainsi, nous fixons les critères suivants sur les problèmes qui seront étudiés. (1) Les problèmes étudiés doivent avoir un espace de recherche étendu pour pouvoir confronter données et prédictions de manière approfondie ; (2) la bonne réponse doit contredire la contrainte d'utilisation de tous les nombres une seule fois. Ces critères sont respectés par les problèmes complexes, exemple ci-après.

> *Au supermarché, le kilo de poisson a augmenté de 5 euros cette année. Un kilo de poisson coûte maintenant 12 euros. Au début de l'année, le kilo de viande coûtait le même prix que le kilo de poisson. Le kilo de viande a augmenté de 3 euros de moins que le kilo de poisson. Combien coûte le kilo de viande maintenant ?*

Ils peuvent être résolus de deux manières différentes, soit en trois étapes, soit en une étape. Une première résolution possible est de déduire par le biais des deux dernières phrases qu'à la fin de l'année le kilo de viande vaut 3 euros de moins que le kilo de poisson. En effet s'il a augmenté de 3 euros de moins que celui du poisson et qu'il avait le même prix au début de l'année alors cet écart de 3 se reporte sur les quantités en fin d'année. La réponse est donc donnée par 12-3=9. Une autre possibilité est de calculer d'un côté le prix de la viande à l'état initial qui correspond au prix du kilo du poisson et qui est donc donné par le calcul 12-5=7. D'un autre côté, l'augmentation du prix de la

viande peut être trouvée par le calcul (5-3). Enfin, le prix final de la viande peut être donné par la somme de ces deux quantités.

Cette double possibilité de résolution a encouragé l'utilisation de ce type de problème dans différentes études. En effet, ils constituent un matériel intéressant pour étudier ce qui, dans les phrases, forme la saillance d'une relation mathématique, l'idée étant que plus une relation est saillante, plus la stratégie qui utilise cette relation est favorisée. Dans notre perspective d'analyse, cette double possibilité de solution est tout aussi intéressante, car la contrainte d'utiliser chaque nombre une seule fois peut être violée de deux manières différentes sans que le résultat soit faux : en réutilisant un nombre du problème (pour la stratégie en 3 étapes) ou en en laissant un de coté (pour la stratégie en une étape). Le formalisme des contraintes nous permet d'analyser cette distinction de manière souple. Désormais, nous parlons de deux contraintes différentes : (1) ne pas réutiliser de nombre et (2) ne pas laisser de nombres de côté. Ainsi la contrainte *« utiliser chaque nombre une et une seule fois »* est simplement l'agrégation de ces deux contraintes. Les problèmes et les données de (Chaillet, 2014) présentés à la suite de notre problématique fournissent donc un matériel de qualité pour tester notre modèle.

#### 8.2.1.2 Quel espace de recherche ?

Entre une théorie et son implémentation dans un modèle informatique, il existe plusieurs couches qui laissent apparaitre des degrés de liberté. Ces choix d'implémentation sont critiques, car ils peuvent mener à des modélisations et prédictions différentes pour une même théorie testée. Ils doivent donc être soigneusement justifiés pour pouvoir conceptualiser les résultats obtenus non pas comme l'évaluation d'une théorie, mais comme l'évaluation d'un modèle qui la représente. Recourir au plus petit nombre d'hypothèses auxiliaires augmente la falsifiabilité du modèle, car s'il est mis en échec, il est « moins facile » de reporter cet échec sur une hypothèse auxiliaire qui pourrait être alors revue de manière ad hoc. Nous explicitons donc les choix de modélisation qui ont été faits avant de présenter les méthodes d'analyse employées et les résultats de ce modèle.

Depuis les travaux fondamentaux de Newell et Simon (1971), représenter la résolution d'un problème par un graphe est devenu une méthode classique de modélisation cognitive. Cette méthode nécessite de faire des choix dans la construction de cet espace. Pour certains problèmes de changement d'état comme pour le problème des tours de Hanoï, les nœuds sont choisis sans ambigüité comme des états physiques de la tour. Il

est difficile d'en faire autant pour la résolution de PAEV. D'ailleurs, dans le problème de la tour de Hanoï, les états illégaux (par exemple lorsqu'un anneau est supporté par un anneau plus petit) ou les états intermédiaires supplémentaires (par exemple, lorsqu'un sujet déplace un anneau et réfléchit où le redéposer) pourraient éventuellement être représentés. L'espace de recherche, rarement neutre vis-à-vis du modèle et des intentions du chercheur, n'est pas une donnée du réel, mais un outil pour formaliser des comportements (ou des modèles du comportement)

Nous proposons donc de construire un espace dans lequel représenter la résolution d'un PAEV. Celui-ci est constitué d'états représentant les quantités calculées et la représentation courante du problème. Il est parcouru par :

- Des opérations cognitives expertes consistant à calculer des quantités inconnues sur la base de quantités connues.
- Des opérations cognitives de réinterprétations consistant à changer le sens de certaines phrases du problème. Cette opération est un relâchement de contrainte.

Ces représentations alternatives sont construites sur la base des travaux de Cummins (1988) et des chercheurs ayant étudié les difficultés linguistiques et conceptuelles de certains éléments de langage. Nous faisons une liste des réinterprétations alternatives choisies pour notre modèle ; nous incluons aussi des inversions liées. Les relâchements des contraintes didactiques liées à l'utilisation des nombres ne sont pas implémentés, car triviaux. Pour ne pas limiter l'exploration, ces contraintes sont considérées comme relâchées dès l'état initial. La forme finale de la réponse générée suffira pour déterminer si ces contraintes ont été relâchées ou pas au cours de la résolution. Nous y reviendrons donc lors de l'analyse des données.

## 8.2.2 Écriture d'un problème

### 8.2.2.1 Création de la structure d'un problème

Nous reprenons l'exemple du problème vu précédemment et détaillons la manière dont il est décrit dans notre programme en précisant les lignes de codes.

> *Au supermarché, le kilo de poisson a augmenté de 5 euros cette année.*
> *Un kilo de poisson coûte maintenant 12 euros.*
> *Au début de l'année, le kilo de viande coûtait le même prix que le kilo de poisson.*

*Le kilo de viande a augmenté de 3 euros de moins que le kilo de poisson. Combien coûte le kilo de viande maintenant ?*

Ces lignes sont écrites sur la base de bibliothèques construites en amont. Le « notebook » électronique d'où sont tirées ces captures peut être récupéré en ligne[31]. Les structures mathématiques et sémantiques du problème sont déclarées avant la construction de la liste de réinterprétations. La structure mathématique produit les objets du problème. Cette construction des problèmes est tout à fait générique et peut être adaptée à des problèmes ayant des structures additives différentes.

Slide Type

```
schema1=Schema("PoissonEF","PoissonEI",operations.addition,"PoissonGAIN","change")
schema2=Schema("ViandeEF","ViandeEI",operations.addition,"ViandeGAIN","change")
print(schema1.objects)
print(schema2.objects)
```

```
{'q1': 'PoissonEI', 'q2': 'PoissonGAIN', 'qf': 'PoissonEF'}
{'q1': 'ViandeEI', 'q2': 'ViandeGAIN', 'qf': 'ViandeEF'}
```

Figure 22. Création des deux schémas principaux constitutifs d'un problème en plusieurs étapes. Affichage des quantités qui le constituent. EI et EF sont des acronymes pour État Final et État Initial.

La première étape est de décrire les relations mathématiques des quantités du problème. La structure du modèle est décrite par des schémas dans le sens classique des travaux de Riley & Greeno (1988). Cette création de schémas, visible en Figure 22, va permettre de décrire de manière sémantique les quantités de l'énoncé pour différencier leurs usages possibles. Le statut des quantités est indiqué par leurs noms, par exemple ViandeEI correspond au prix de la viande dans l'état initial. Les quantités ne sont pas toujours présentées dans le texte, mais dans la mesure où elles peuvent être utilisées pour résoudre le problème, elles doivent être disponibles et inscrites dans les relations mathématiques qui les définissent. C'est le sens et le rôle de la fonction « addBridgingSchemas » visible en Figure 23.

[31] L'ensemble des ressources de ce type (notebooks, programmes et analyses statistiques) est disponible par le biais de liens donnés en annexe.

```
struct=ProblemStructure()
struct.addSchema(schema1)
struct.addSchema(schema2)
struct.addBridgingSchemas(schema1,schema2) # fonction qui ajoute les schémas suivant
struct.updateObjectSet() # fonction qui mets à jours les nouvelles quantités du schéma

# affichage des nouveaux objets de l'énoncé
ls=struct.schemas
for schem in ls:
    print(schem.objects)
```

```
{'q1': 'PoissonEI', 'q2': 'PoissonGAIN', 'qf': 'PoissonEF'}
{'q1': 'ViandeEI', 'q2': 'ViandeGAIN', 'qf': 'ViandeEF'}
{'q1': 'PoissonEI', 'q2': 'ViandeEI', 'qf': 'PoissonEIminusViandeEI'}
{'q1': 'PoissonGAIN', 'q2': 'ViandeGAIN', 'qf': 'PoissonGAINminusViandeGAIN'}
{'q1': 'PoissonEF', 'q2': 'ViandeEF', 'qf': 'PoissonEFminusViandeEF'}
{'q1': 'PoissonEIminusViandeEI', 'q2': 'PoissonGAINminusViandeGAIN', 'qf': 'PoissonEFminusViandeEF'}
```

Figure 23. Finalisation de l'écriture de la structure du problème. Affichage de l'ensemble des éléments des schémas du problème.

Cette fonction vient compléter l'ensemble de ces variables, en produisant toutes les correspondances possibles entre les deux schémas du problème. Quatre nouveaux schémas sont donc produits :

- Un schéma décrivant les relations entre les nouvelles quantités produites (en gris dans la partie droite de la Figure 24).
- Les trois nouveaux schémas directement liés aux correspondances produites (en couleurs dans la partie droite de la Figure 24).

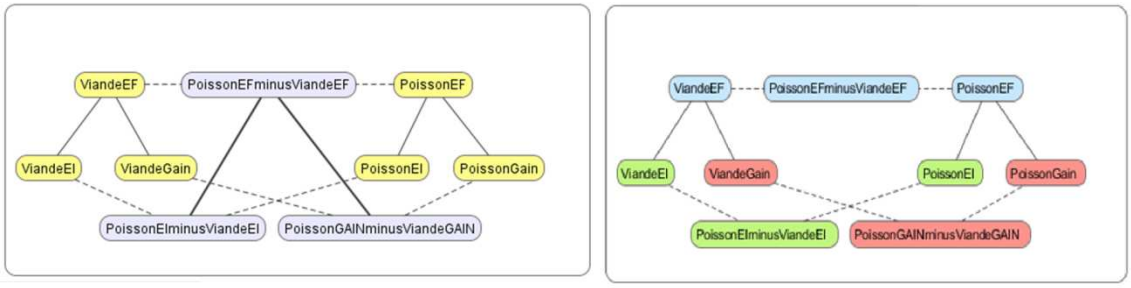



Figure 24. Schémas supplémentaires produits par la fonction « addBridgingSchemas ». Les couleurs et les liens permettent d'indiquer les appartenances aux différents schémas du problème.

8.2.2.2 Création de la partie sémantique du problème

La Figure 25 présente la manière dont les propositions du problème sont décrites. Nous évitons la problématique du traitement du langage naturel en représentant les informations du texte comme des déclarations de quantités. Ces dernières sont conçues comme l'association d'une référence à un nombre du texte et à un objet du schéma. La

première ligne du problème est donc l'association entre l'objet « PoissonGain » et le nombre P1[32].

```
t1='Au supermarché, le kilo de poisson a augmenté de 5 euros cette année'
t2='Un kilo de poisson coute maintenant 12 euros.'
t3='Au début de l\'année, le kilo de viande coutait le même prix que le kilo de poisson.'
t4='Le kilo de viande a augmenté de 3 euros de moins que le kilo de poisson'
q1=Representation(Quantity("PoissonGAIN","P1"),t1)
q2=Representation(Quantity("PoissonEF","T1"),t2)
q3=Representation(Quantity("PoissonEIminusViandeEI","dEI"),t3)
q4=Representation(Quantity("PoissonGAINminusViandeGAIN","d"),t4)
text.addTextInformation(TextInformation(q1))
text.addTextInformation(TextInformation(q2))
text.addTextInformation(TextInformation(q3))
text.addTextInformation(TextInformation(q4))
```

Figure 25. Écriture de la partie sémantique du problème

Nous n'abordons pas en détail la construction du modèle de l'élève. Un « solveur » est décrit par deux composantes.

- Un vecteur de ses représentations des propositions du problème composé à l'état initial de 0 (0,0,0,0) pour signifier que ce sont les représentations expertes des propositions de l'énoncé qui sont activées. Lorsqu'une interprétation alternative est adoptée, alors l'index de la représentation sélectionnée est noté. Par exemple (0,0,2,0) signifie que le « solveur » a adopté la deuxième représentation alternative pour la troisième proposition du problème.
- Un dictionnaire des quantités qu'il a à sa disposition. Elles peuvent être issues :
    - Des quantités calculées.
    - De sa représentation du problème (représentée par le vecteur plus haut).

Lorsque ce dictionnaire contient la quantité demandée dans la question, le problème est résolu.

[32] Nous avons choisi de travailler avec les références des nombres plutôt qu'avec les nombres directement pour conserver un niveau de généralité « haut » dans la création de problèmes. Une fonction, non représentée ici, permet donc d'instancier les nombres du problème.

#### 8.2.2.3 Écriture des réinterprétations

Chacune des propositions du problème a une interprétation experte, c'est celle que nous avons décrite dans le paragraphe plus haut. Chacune a aussi plusieurs interprétations alternatives sur la base desquelles la résolution aura lieu. La Figure 26 montre comment cette étape est réalisée.

```
t1='Au supermarché, le kilo de poisson était de 5 euros'
t2='Au supermarché, le kilo de poisson coute 5 euros'
t3='Un kilo de poisson était de 12 euros.'
t4='Au la fin de l\'année, le kilo de viande coute le même prix que le kilo de poisson.'
t5='Le kilo de viande a augmenté du même prix que le kilo de poisson.'
t6='Le kilo de viande a augmenté de 3 euros'
t7='Le kilo de viande a diminué de 3 euros'
t8='Le kilo de viande vaut 3 euros de moins que le kilo de poisson'
t9='Le kilo coute 3 euros à la fin'

a1=Representation(Quantity("PoissonEI","P1"),t1)
a2=Representation(Quantity("PoissonEF","P1"),t2)
a3=Representation(Quantity("PoissonEI","T1"),t3)
a4=Representation(Quantity("PoissonEFminusViandeEF","dEI"),t4)
a5=Representation(Quantity("PoissonGAINminusViandeGAIN","dEI"),t5)
a6=Representation(Quantity("ViandeGAIN","d"),t6)
a7=Representation(Quantity("ViandeGAIN","-d"),t7)
a8=Representation(Quantity("PoissonEFminusViandeEF","d"),t8)
a9=Representation(Quantity("ViandeEF","d"),t9)

text.getTextInformation(0).addAlternativeRepresentation(a1)
text.getTextInformation(0).addAlternativeRepresentation(a2)
text.getTextInformation(1).addAlternativeRepresentation(a3)
text.getTextInformation(2).addAlternativeRepresentation(a4)
text.getTextInformation(2).addAlternativeRepresentation(a5)
text.getTextInformation(3).addAlternativeRepresentation(a6)
text.getTextInformation(3).addAlternativeRepresentation(a7)
text.getTextInformation(3).addAlternativeRepresentation(a8)
text.getTextInformation(3).addAlternativeRepresentation(a9)
```

Figure 26. Ajout des représentations alternatives aux propositions du problème.

Voici, dans un format plus lisible, l'ensemble des réinterprétations que nous avons proposé pour le problème Tc4t.

Au supermarché, le kilo de poisson a augmenté de 5 euros cette année

- Au supermarché, le kilo de poisson était de 5 euros
- Au supermarché, le kilo de poisson coûte 5 euros

Au début de l'année, le kilo de viande coûtait le même prix que le kilo de poisson.

- Un kilo de poisson était de 12 euros

Le kilo de viande a augmenté de 3 euros de moins que le kilo de poisson

- Le kilo de viande a augmenté de 3 euros
- Le kilo de viande a diminué de 3 euros

- Le kilo de viande a augmenté de 3 euros de plus que le kilo de poisson
- Le kilo de viande vaut 3 euros de moins que le kilo de poisson
- Le kilo de viande vaut 3 euros de plus que le kilo de poisson
- Le kilo coûte 3 euros à la fin

Les problèmes ont des représentations alternatives qui se recoupent. Pour faciliter leurs écritures, nous utilisons une fonction de copie pour faciliter la tâche. Par exemple dans un autre problème (Tc1t) pour lequel le contexte sémantique change un peu, seule une réinterprétation est ajoutée.

Ainsi, la phrase « Après la récréation, Lucas a 16 billes. » peut être réinterprétée comme « Avant la récréation, Lucas avait 16 billes. » ainsi que comme cette proposition supplémentaire : « Après la récréation, Lucas gagne 16 billes » qui repose sur la proximité de sens entre le verbe « avoir » et le verbe « obtenir ».

Cette opération est proposée pour d'autres problèmes ; par exemple, « Après janvier, il y a 16 enfants à la chorale », qui peut être interprétée comme un gain. Un autre type de réinterprétation dépendant du contexte sémantique peut être signalée dans le problème « Cc2t ». À la phrase « *Quand Médor et Rex montent ensemble sur la balance chez le vétérinaire, la balance indique 15 kilos.* » Les réinterprétations « *Médor pèse 15 kilos* » ainsi que « *Rex pèse 15 kilos* » sont alors associées à cette phrase.

## 8.2.3 Les choix de conception

### 8.2.3.1 Choix d'homogénéisation : calcul de quantité comme seul type d'inférence valide.

Des choix ont été faits pour homogénéiser le modèle. Aucune quantité n'est définie comme la composition d'autres quantités, par exemple la différence entre le prix de la viande et du poisson est représentée par un objet qui lui est propre « d ». Cela permet de modéliser la compréhension de « le prix de la viande est 3 euros plus élevé que le prix du poisson » comme l'association d'une quantité à un objet tel que nous l'avons expliqué plus haut. Cela vaut aussi pour les quantités de plus hauts degrés comme des différences de différences. Toutes les variables dont le calcul peut représenter un pas dans la résolution sont donc prédéfinies. Le calcul de ces quantités décrivant des comparaisons est rendu possible par la fonction addBridgingSchemas précédemment.

**Inférence comme calcul de quantité.** Pour la fluidité de la modélisation, il est important de conserver l'idée que toute interprétation de proposition se traduit comme l'association d'un objet avec une quantité. Or, il peut être nécessaire pour l'enfant d'utiliser une égalité qui est décrite dans le texte. Dans le problème Tc4t par exemple, il est indiqué que le prix de la viande égale le prix du poisson à l'état initial. Pour cette raison, lorsque le problème décrit une égalité, nous la considérons comme une relation entre deux quantités en faisant intervenir un 0. Deux opérations faisant intervenir un 0 sont donc possibles :

- Utiliser le fait que le prix de la viande égale celui du poisson consiste à faire une addition par 0.
- L'inférence dite de *« la partie commune »* permet de reporter un écart lorsque deux schémas ont une partie commune. Nous l'avons décrite lors de la présentation du problème Tc4t. Cette inférence est aussi représentée comme un calcul faisant intervenir un 0.

Si d'un point de vue cognitif, réaliser une inférence et une opération « *nulle* » n'ont pas le même sens, le projet de modélisation présenté ici n'a pas de raison de s'alourdir de cette distinction. Dans le cadre d'un modèle cognitif où la difficulté relative des opérations cognitives est prise en compte, il serait certainement opportun de distinguer les opérations des inférences (simples et complexes).

#### 8.2.3.2 Utilisation des réinterprétations

Nous avons choisi de ne pas implémenter de stratégie « intelligente » de résolution, préférant l'exploration combinatoire de toutes les possibilités de résolutions, composée de réinterprétations et de calculs experts. Se pose alors la question de la limitation du nombre de réinterprétations produites par le modèle. Augmenter cette limite fait croître le nombre de prédictions. Si l'hypothèse de la présence de réinterprétations est faite, moins elles sont nombreuses, plus le modèle a de chance de voir ses prédictions confirmées. En effet, si nos hypothèses sont justes, les réponses d'élèves passant par deux réinterprétations sont plus rares que celles en ayant une seule. Dès lors, si le modèle à une interprétation produit suffisamment de prédictions, il est préférable de centrer nos analyses sur ce premier modèle avant d'étendre à deux réinterprétations si ses résultats sont positifs.

Un autre questionnement porte sur le moment où ces réinterprétations peuvent avoir lieu. Il est possible de ne considérer que les réinterprétations qui ont lieu au départ de la

résolution, c'est à dire avant tout calcul. Nous testerons ce modèle qui portera le nom de modèle « direct ».

8.2.3.3 Prise en compte du niveau d'abstraction d'une variable

Les variables de premier degré (les billes de pierre, la taille de Jean) sont différenciées des variables de second degré exprimant des comparaisons entre ces grandeurs (la différence entre le nombre de billes de Pierre et de Jean) et de troisième degré (la différence entre l'augmentation de prix du poisson et celui de la viande au marché cette année). Les variables de premier degré ne peuvent pas être négatives, alors que celles de degré plus haut peuvent l'être, car elles expriment des comparaisons. Ainsi nous ajoutons des contraintes à l'exploration des chemins possible par une fonction dédiée à cette tâche présentée en Figure 27.

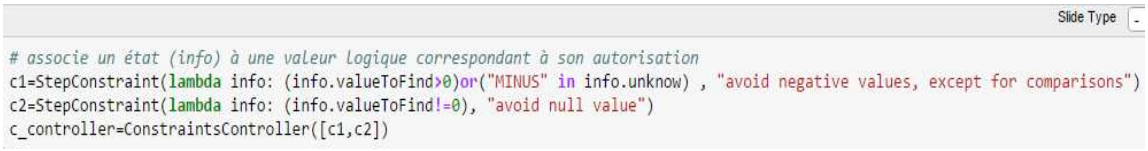



Figure 27. Création de contraintes pour limiter certains états du problème. En première ligne, seules les valeurs positives sont autorisées sauf si la quantité est une comparaison (elle contient le terme « MINUS »). En deuxième ligne, les valeurs nulles sont interdites, quel que soit le type de quantité.

8.2.3.4 Gestion des incohérences par le principe d'exploration maximale

Un dernier ensemble de règles porte sur la gestion des incohérences. Elles apparaissent lorsqu'une quantité prend une valeur en désaccord avec une ancienne valeur. Le fait que l'interprétation des propositions varie au cours de la résolution rend cette situation fréquente. Nous avons choisi une fois de plus de prendre le parti de l'exploration en favorisant la création de nouvelles quantités. De ce fait, si une réinterprétation vient modifier une quantité calculée plus tôt, elle l'écrase.

Une autre question difficile porte sur la coexistence des réinterprétations avec le sens originel de la phrase. Les quantités appartenant à l'ancienne interprétation sont laissées disponibles. Autrement dit, même après avoir réinterprété une phrase, son sens originel reste disponible. Un dernier cas représentant une incohérence est la présence d'une partie plus grande qu'un tout. Par exemple, la collection des billes de Thomas ne peut pas être plus grande que celles de Thomas et Jean réunis. Toujours sous l'idée de favoriser l'exploration, ces états n'ont pas été interdits, mais le calcul de la partie restante a, lui, été interdit pour éviter de trouver des nombres négatifs.

Des modélisations ultérieures pourraient gérer les incohérences par d'autres contraintes. Par défaut, les incohérences sont évitées, mais les contraintes représentant ce système de cohérence peuvent être relâchées. Une telle modélisation permettrait par exemple de rechercher si ces contraintes sont globalement respectées et si certains élèves y accordent un poids plus fort.

#### 8.2.3.5 Synthèse de l'approche

Ce modèle peut être reformulé comme la construction et l'exploration d'un espace de recherche dont les nœuds sont représentés par (1) l'ensemble des quantités dont le sujet dispose (celles calculées et celles de l'énoncé), (2) l'état des interprétations des propositions de l'énoncé et (3) le nombre de réinterprétations déjà effectuées. Un état final est un état dans lequel la quantité demandée dans l'énoncé est trouvée. Les déplacements dans cet espace sont composés du calcul de quantités, d'inférences et de réinterprétations de phrases de l'énoncé. Ils détiennent les contraintes que nous résumons ici :

- Pas de variables inférieures ou égales à 0 (sauf issues de comparaison qui peuvent être négatives)
- Une seule réinterprétation au cours de la résolution.

#### 8.2.3.6 Synthétise des prédictions établies par le modèle

Pour illustrer les règles et les prédictions du modèle, nous allons récupérer celles qui sont construites à partir de la séquence suivante : (1) le modèle réinterprète une phrase, (2) il cherche à résoudre le problème en utilisant les règles de calculs expertes de quantité.[33]

Par explosion combinatoire, le modèle produit une grande quantité de voies menant à la réponse de chaque problème. Le problème présenté, par exemple, génère plus de 104 chemins possibles respectant la séquence.

[33] Nous rappelons que pour établir l'ensemble des prédictions, la phase de réinterprétation peut avoir lieu n'importe quand dans la résolution. Nous présentons donc ici une utilisation simplifiée du modèle en guise d'exemple.


Figure 28. Exploration de l'arbre des solutions possibles construit par l'objet solver. Seules les trois premières possibilités sont affichées. Des traits bleus ont été rajoutés pour mettre en évidence la structure arborescente.

Une fois que le « solveur » a été exécuté et a exploré les chemins possibles (appel des fonctions non représentées ici), il est possible d'afficher les chemins découverts sous la forme d'arbre représenté en Figure 28.

Le nombre élevé de chemins possibles résulte du fait que beaucoup de quantités inutiles peuvent être utilisées et que l'ordre de beaucoup de calculs peut être interverti. En regardant la Figure 28, il apparait que les deux chemins mènent à la réponse 7 en passant par la réinterprétation « *5 interpreted as poisson EI* ». Elle correspond à l'interprétation de la phrase *au supermarché, le kilo de poisson a augmenté de 5 euros cette année* par *au supermarché, le kilo de poisson* ***était*** *de 5 euros au début de l'année.* À partir de cette réinterprétation la réponse (erronée) 7 peut être obtenue en déduisant que :

- La viande valait aussi 5 euros cette année (par l'utilisation de la phrase « *Au début de l'année, le kilo de viande coûtait le même prix que le kilo de poisson* »). Cette inférence est représentée comme un calcul faisant intervenir un zéro.

- La viande a augmenté de 5-3=2 euros en utilisant « Le kilo de viande a augmenté de 3 euros de moins que le kilo de poisson. » et « le kilo de poisson a augmenté de 5 euros cette année. » [34]
- La viande vaut 5+2=7 euros à l'état final en utilisant les deux données calculées précédemment.

L'algorithme de résolution produit parfois des calculs de quantité inutiles, visibles en comparant les deux premières solutions représentées dans l'arbre en Figure 28. La seule différence avec la branche présentée est l'inférence 0+3=3 trouvant l'objet « PoissonEFMinusViandeEF » n'ayant aucune utilité pour la suite des calculs. Ces calculs, générés fréquemment par le modèle, nous intéressent peu, mais résultent de notre choix d'implémentation sans planification.

Pour limiter la redondance des prédictions, les propositions du modèle sont résumées par deux éléments :

- Les réinterprétations produites (et effectivement utilisées)
- La formule globale représentant l'ensemble des calculs effectivement utilisés sur une seule ligne (e.g (T1-P1)-(P1+d)) et respectant les mêmes règles de priorités dans la notation élève (le plus grand nombre en premier opérande et le premier calcul le plus à gauche).

Les 104 productions peuvent ainsi être réduites à 22, il n'est pas pertinent de noter combien de fois une formule globale est prédite car les aspects quantitatifs n'ont pas de sens ici, nous notons simplement qu'elle a été prédite. L'ensemble des prédictions établies pour le problème de cette partie est donné en annexe.

[34] Conformément au passage sur la gestion des incohérences, il est possible d'utiliser « le kilo de poisson a augmenté de 5 euros cette année. » même si cette phrase a été réinterprétée.

### 8.2.4 Construction du modèle des mots-clefs

Pour affiner nos réflexions et parce que nous avons remis en cause la théorie de la résolution par mots-clefs dans notre revue de la littérature, nous avons souhaité comparer notre modèle avec un modèle qui incarnerait ce type de résolution.

Mettre en place un modèle de résolution par mots-clefs n'est pas aussi direct pour les problèmes complexes que pour les problèmes présentant simplement deux nombres. Peu de questions se posent pour ces problèmes, il suffit de prendre l'opération associée au mot-clef le plus saillant. « *Marie a des billes, elle en perd 9, maintenant elle en a 4. Combien en avait-elle avant ?* » Par la présence du mot « perd », la résolution par mots-clefs donne la réponse erronée *9-4=5.*

À notre connaissance, la littérature mentionnant l'existence de cette stratégie n'établit pas de description suffisamment précise pour l'implémenter directement sur des problèmes plus complexes. Dans ces problèmes, il y a parfois plusieurs mots-clefs, parfois conflictuels. En extrayant les nombres et les mots qui donnent des indices sur les opérations du problème précédent, nous obtenons le Tableau 10.

Tableau 10. Extraction des mots-clefs d'un problème complexe.

| | |
|---|---|
| Au supermarché, le kilo de poisson a **augmenté** de **5** euros cette année. | +5 |
| Un kilo de poisson coûte maintenant **12** euros. | 12 |
| Au début de l'année, le kilo de viande coûtait le même prix que le kilo de poisson. | rien |
| Le kilo de viande a **augmenté** de **3** euros de **moins** que le kilo de poisson. | + 3 - |
| Combien coûte le kilo de viande maintenant ? | rien |

Si 5+12 et 12+3 semblent les prédictions les plus directes qu'un modèle de mots-clefs puisse établir, il est pensable d'inclure des formules en plusieurs étapes comme (5+12) +/-3 et (12-3) + 5 ou encore 12-3 et des formules réunissant des nombres éloignés comme 5+/-3. Tout comme le modèle de réinterprétation, il est important de produire suffisamment de prédictions pour pouvoir augmenter la puissance statistique des analyses. C'est pourquoi nous favoriserons ici aussi une logique d'exploration maximum en évitant tout de même (1) de produire trop de prédictions, (2) de faire des prédictions qui contrediraient une logique par mots-clefs.

Il a donc été choisi de fonctionner par la négative : par élimination des formules qui ne s'y apparentent pas. Voici les règles proposées pour savoir si une formule est ou n'est pas en contradiction avec une approche par mots-clefs.

- Interdiction de toute formule qui utilise le signe opposé associé à un nombre.

Ici : 12-5 est donc interdit, ainsi que toute expression du type 5-(calcul) et (calcul)-5.

- Les nombres qui n'ont pas d'indice associé peuvent être utilisés dans des formules, mais pas plus d'une fois.

Ici : (12+5)+12 n'est pas possible.

- Si deux éléments n'ont pas de mots-clefs, alors nous interdisons les opérations qui prennent ces deux nombres en opérandes.

Ici : Le cas n'est pas présent. Pour le besoin de l'exemple, si 5 n'avait pas d'indice, alors 5+12 et 12-5 auraient été interdits.

- Contrainte d'opération. Si x1 est le nombre qui est lié à un indice de soustraction, alors x2-x1 est autorisée, mais x1-x2 ne l'est pas.

L'idée sous-jacente est que « 3 de moins », « 5 de plus » peuvent être vus comme des opérateurs qu'on applique à un nombre existant. Cette règle peut être débattue, mais d'un point de vue pratique, dans les problèmes que nous étudions, cette règle n'a que peu d'effet, car c'est toujours « d » qui est associé à un « moins », et d est généralement petit, donc le cas est très rare. Pour l'addition, cela n'a pas de conséquence, car les prédictions du modèle des réinterprétations ne portent pas sur l'ordre des opérandes.

Ici : Pas de 3-....

Toutes les opérations qui ne sont pas interdites par ces règles sont donc autorisées. Voici la liste des patterns interdits pour Tc4t : « -5, 5-, 3- , (12+5)+12».

Toutes les autres formules sont susceptibles d'être interprétées comme la résultante d'une résolution basée sur les mots-clefs.

Choisir quels mots-clefs sont des indices d'opération est une étape délicate. Nous avons donc constitué deux modèles de mots-clefs, l'un utilise une liste restreinte avec des associations plutôt consensuelles, l'autre une liste étendue avec des mots-clefs clefs plus discutables. Nous testerons donc deux modèles de mots-clefs selon que l'extension est sélectionnée ou non.

- Liste consensuelle : « augment », « moins », « plus », « diminu », « gagne »
- Liste faisant extension : « pris », « ensemble », « après »,« avant », « réuni »

Ces éléments sont directement recherchés dans les phrases du problème, d'où la présence de mots-clefs « tronqués ». « augment » pourra donc être détecté pour les mots-clefs « augmenter », « augmentation » etc.

## 8.3 Matériel

### 8.3.1 Adaptation des données fournies

Nous reprenons les données présentées à la suite de la problématique. À ce stade une information supplémentaire est à prendre en compte. Deux expérimentations sur des CM1 et CM2 ont été réalisées. Dans la deuxième expérimentation, de légères variations ont été mises en place dans les textes des problèmes et dans les choix des nombres. Bien que la plupart des problèmes laissent inchangées les prédictions (changement de place de phrase et de nombres), certains (les problèmes d'argent) ont été reformulés plus largement. Compte tenu des variations, nous avons choisi de ne pas fusionner les prédictions sur les problèmes, nous considérons donc l'ensemble des 32 problèmes posés. Les 16 problèmes de ces deux expérimentations sont donnés en annexe.

Nous enlevons (dans les prédictions des modèles et dans les données) les bonnes réponses pour deux raisons. Elles sont évidemment fréquentes et appartiennent au set de prédiction du modèle, elles sont obtenues sans réinterprétation. De ce fait, les garder conduirait à un biais considérable en faveur du modèle, car ce dernier n'a pas été établi pour prédire de bonnes réponses, mais des erreurs. Ses qualités statistiques seraient artificiellement améliorées par ce manque de spécificité dans l'analyse. Nous supprimons également les absences de réponses et les erreurs qui sortent de notre champ d'analyse (comme les quelques cas d'utilisation de multiplication ou de divisions).

### 8.3.2 Méthode d'analyses

Les résultats ne peuvent être analysés de manière directe et simple. L'établissement des p-values et des tailles d'effet suit une méthodologie qu'il convient de détailler.

#### 8.3.2.1 Construction de l'espace de réponses *a priori*

Avant toute démarche de comparaison, il est important de définir le terrain commun sur lequel les données et les prédictions vont être comparées. Nous nommons le générateur

de ces formules *a priori* possibles GenAv (pour **Gen**érateur **Av**eugle). Nous verrons dans les analyses que les prédictions comptent autant que les non-prédictions. L'approche que nous avons choisie compare la popularité des réponses prédites comparées aux réponses non prédites. De ce fait, la construction de cet espace de formules possibles est une étape sensible dans la mesure où la considération de réponses complètement irréalistes augmente artificiellement les qualités statistiques du modèle si celles-ci ne sont pas prédites. De manière symétrique, et comme nous l'avons indiqué précédemment, inclure dans cet espace les bonnes réponses aux différents problèmes causerait un biais important en faveur du modèle. En effet les bonnes réponses sont relativement fréquentes et le modèle prédit évidemment ces bonnes réponses, mais elles ne constituent pas un argument particulièrement en faveur du modèle. Les règles que respecte GenAv sont les suivantes :

- **Limitation du nombre de calculs**. Pour éviter la production de réponses trop peu fréquentes, le nombre de calculs (effectivement utilisés) est limité à trois. En effet, les élèves enchaînent très rarement plus de 3 calculs pour des problèmes qui ne contiennent que trois nombres.
- **Évitement des nombres négatifs.** Une autre contrainte consiste à ne pas produire de suite de calculs aboutissant sur un nombre négatif. Il est vrai que notre modèle peut produire des nombres négatifs dans certains cas (comme expliqué dans la partie précédente), mais si GenAv est autorisé à produire des nombres négatifs, le nombre total de prédictions seraient presque multipliées par deux, ce qui est une inflation considérable. Cette règle sur l'absence de valeurs négatives ou égales à 0 implique de légères variations sur l'ensemble des formules générées selon les différents problèmes. Par exemple, la réponse (T1-P1)-(P1+d) n'est possible que si l'inégalité T1>2P1+d est vérifiée ce qui n'est pas le cas dans tous les problèmes[35].

[35] Inégalité vérifiée pour les 6 problèmes : Cc2p, Tc2p, Cc3p, Cc2t, Cc3t, Tc2t

- **Évitement de certaines opérations**. Toujours pour limiter l'artificialité des réponses construites par GenAv, nous ne considérons pas les calculs du type T1+T1 ou P1-P1. Ces calculs sont évidemment rares et ne sont pas prédits par le modèle. Pour les raisons décrites plus haut, les bonnes réponses sont aussi soustraites de cet espace de réponse possible.

Après la restriction des réponses à cet espace d'analyse, il reste 1904 erreurs réelles d'élèves qui attendent d'être expliquées par nos travaux de modélisation. Les types d'erreurs possibles générés par GenAv varient selon les 32 problèmes. Pour le problème Tc4t par exemple, 184 types d'erreurs possibles sont générés. Au total, c'est-à-dire, en cumulant ces types d'erreurs au travers l'ensemble des problèmes, 5893 erreurs sont prévues par GenAv.

#### 8.3.2.2 Analyse brute des résultats

Avant d'introduire les raisons pour lesquelles nous avons choisi de mener nos analyses par des tests de permutation stratifiés, nous introduisons la version la plus simple des analyses et nous détaillons ses limites. Ce n'est donc pas à proprement parler une présentation des résultats, mais une partie permettant d'introduire les méthodes d'analyse proposées.

Nous rappelons que les prédictions du modèle sont binaires, car certains types de réponses sont prédites et d'autres ne le sont pas. Les données, au contraire, sont quantitatives. À chaque type d'erreur produit par GenAv, nous pouvons y associer le nombre de fois qu'elle a été observée dans les données.

Cette analyse se base sur le report de ces deux valeurs :

- Le pourcentage de prédictions sur l'espace des possibles. Cet espace est construit par GenAv.
- Le pourcentage d'occurrences dans les données qui sont effectivement prédites par le modèle. Nous parlons alors de pourcentage de réponses capturées.

Intuitivement, il est souhaitable que :

- La première valeur soit la plus petite possible : l'aire des prédictions du modèle est faible relativement à l'aire maximale des prédictions possibles.

- La deuxième valeur soit la plus grande possible (le volume de données expliquées par le modèle est grand relativement au volume maximal constitué des 1904 erreurs extraites des données).

Dans le modèle de réinterprétation de base, le pourcentage de prédictions est **de 3.6 %** (211 prédictions sur 5893 possibles) et le pourcentage de capture est **de 56 %** (1071 erreurs capturées par les prédictions sur 1904**).** En d'autres termes, lorsqu'un élève répond de manière erronée, il existe plus de 5 chances sur 10 que la réponse soit prédite par le modèle, alors que les prédictions du modèle ne recouvrent que moins de 4 % de l'ensemble des prédictions possibles. Bien qu'encourageante, cette manière de présenter les résultats n'est pas vraiment correcte, comme nous le verrons dans la partie suivante.

8.3.2.3 Nécessité d'une analyse par strates

Cette mise en correspondance entre prédictions du modèle et données empiriques peut être critiquée, nous expliquons pourquoi. Comparons ces formules de différentes formes :

- P1+d
- T1-(P1+d)
- P1+(P1+d). Ce type de formule diffère de la précédente du fait qu'il y a réutilisation de nombre.
- (T1-P1)-(P1+d)

Nous remarquons qu'elles ont une typologie différente. La répartition des données empiriques ainsi que les prédictions du modèle varient beaucoup selon cette typologie. Or ces variations n'entretiennent pas le même rapport. En effet l'immense majorité des réponses ont la forme de la formule 1 avec 1031 réponses ou la forme de la formule 2 avec 738 réponses contre 1904 erreurs au total. Les deux autres formes sont plus marginales. Or pour ces deux premières formes, le pourcentage de prédiction est plus grand (37/160 pour la forme 1, 96/384 pour la forme 2).

Catégoriser les analyses suivant la forme à laquelle s'apparentent les formules suit un double objectif, le premier est la réduction d'un biais causé par la non-indépendance entre les formes de formules et les valeurs que l'on calcule (pourcentage de recouvrement et pourcentage de capture). Le deuxième objectif est de faire intervenir les réflexions que nous avons amenées dans la partie théorique sur les relâchements des

contraintes liées aux nombres. Si nous reprenons nos contraintes sur ce domaine, nous avions :

- C1 : Utiliser tous les nombres de l'énoncé
- C2 : Ne pas utiliser deux fois un nombre de l'énoncé.

Ces deux contraintes permettent de décrire les quatre catégories de réponse que nous mettons en place selon qu'elles sont respectées ou relâchées. Ainsi les prédictions du modèle peuvent être réparties dans ces quatre colonnes. Le Tableau 11 donne cette répartition pour le problème Tc4t.

Tableau 11. Répartition des prédictions établies pour le problème Tc4t suivant le respect de deux contraintes portant sur l'emploi des nombres.

| | Utiliser tous les nombres de l'énoncé | | |
|---|---|---|---|
| Ne pas utiliser deux fois un nombre de l'énoncé | | | |
| T1+d | T1+(P1+d) | ((T1-d)+P1)+d | (T1-d)-d |
| T1+P1 | (T1+d)+P1 | T1-((P1+d)-d) | (T1+d)-d |
| P1+d | (T1+P1)-d | ((T1+d)+P1)+d | (T1+P1)+P1 |
| | (T1-d)+P1 | T1+((P1+d)+d) | (T1+d)+d |
| | (T1+P1)+d | T1+((P1+d)-d) | (P1+d)+P1 |
| | T1-(P1+d) | (T1+P1)+(P1+d) | (T1-d)+d |
| | | ((T1+P1)+d)+d | (P1+d)+d |
| | | ((P1+d)+P1)+T1 | (P1+d)-d |
| | | .... | |
| | | (28 autres prédictions) | |

Par la suite, nous utiliserons des codes pour distinguer ces quatre catégories de réponse :

- 1C : « un calcul ».
    - Exemple : T1+P1
    - C1 est respectée et C2 est relâchée.

- 2C pour « deux calculs ».
    - Exemple : (T1+P1)-d.
    - Les deux contraintes sont respectées.
- 2C1R pour « deux calculs et une réutilisation de nombre ».
    - Exemple : (T1+P1)-T1.
    - Les deux contraintes sont relâchées.
- 3C pour « trois calculs ».
    - Exemple : T1-((T1-P1)-d).
    - C1 est relâchée et C2 est respectée.

**Note** : Il y a une réutilisation de nombre dans la catégorie 3C, mais comme il est impossible d'avoir une solution en trois calculs sans réutilisation de nombre, il est inutile de le mentionner dans le code de cette catégorie.

8.3.2.4 Choix d'un test de permutation pour tester le modèle

Deux manières de procéder semblent être possibles pour tester le modèle contre le hasard. La première consiste à poser l'hypothèse nulle sur les réponses du sujet *« les réponses se répartissent de manière uniforme »* sur l'ensemble des prédictions possibles produites par GenAv. Ainsi, la probabilité que le modèle capture une réponse d'élève est égale à son recouvrement relatif de l'ensemble des réponses possibles : si le modèle prédit 4 types d'erreurs sur 10, alors il devrait capturer en moyenne, sous l'hypothèse nulle, 4 réponses sur 10, avec un écart-type calculable (2/10 en l'occurrence). Sous l'hypothèse d'uniformité, la probabilité de prédire au moins Nc réponses parmi Ng réponses empiriques par l'utilisation de Pc prédictions parmi Pg prédictions possible est calculable par la loi binomiale suivante :

$$\text{p} - \text{value} = \sum_{k=Nc}^{Ng} \binom{Ng}{k} p^k (1-p)^{Ng-k}$$

avec $p = Pc/Pg$

Cependant, cette formulation de l'hypothèse nulle ne nous semble pas adéquate, car l'idée même que les réponses des sujets se répartissent de manière uniforme est critiquable.

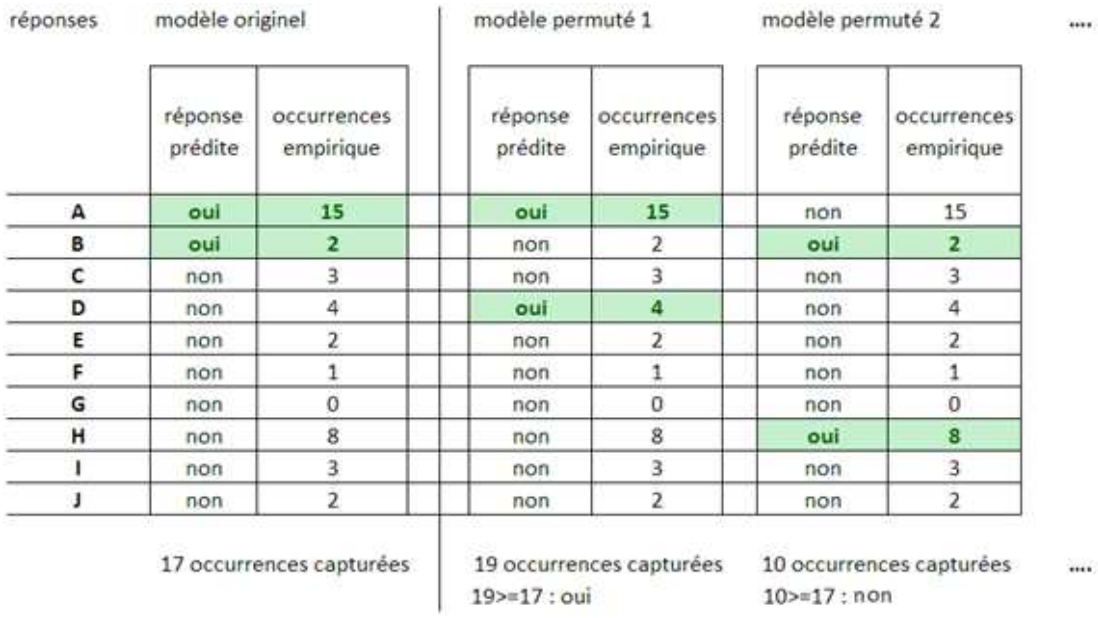
La deuxième possibilité, choisie, considère que sous l'hypothèse nulle le modèle a le même pouvoir explicatif qu'un modèle qui **tirerait au hasard ces prédictions**. Ainsi, la p-value sous cette conception est la probabilité que ce tirage aléatoire réussisse à capturer au moins autant de réponses que la sélection originale du modèle en faisant autant de prédictions. Pour effectuer ce tirage, il suffit d'appliquer une permutation des prédictions que fait le modèle (cf. Figure 29). Dans nos analyses, nous effectuons à chaque fois 100 000 permutations aléatoires. À chaque permutation, le nombre d'occurrences capturées par le modèle aléatoire est enregistré pour former ce qui est appelé une distribution des permutations (« permutation distribution ») de notre statistique d'intérêt (Gibbons & Chakraborti, 2003, p. 288).

Le ratio de modèles aléatoires obtenant des résultats aussi bons ou meilleurs que notre modèle est déterminé. Ce ratio est équivalent à une p-value. Avec cette technique, **nous ne construisons pas l'hypothèse nulle sur la distribution des réponses des sujets, mais nous la construisons sur la distribution des prédictions**.

Cette technique présente l'inconvénient de recourir à une génération aléatoire de données, ce qui limite la rapidité des analyses,[36] mais elle nous semble ici incontournable, car elle permet de ne pas altérer la distribution des réponses des sujets et de rester proche de ce que le modèle fait : produire des prédictions.

Nous présentons en annexe des explications supplémentaires et un exemple mettant en évidence une différence importante sur les p-values obtenues par les deux méthodes lorsque la distribution des réponses n'est pas uniforme. Le test paramétrique donne alors une p-value que nous pouvons considérer comme biaisée.

[36] À titre indicatif, le programme effectuant l'intégralité des analyses présentées dans ce chapitre a mis plus de 6 h à s'exécuter sur un ordinateur aux performances décentes (4GB RAM, 64 bits).



| réponses | modèle originel | | modèle permuté 1 | | modèle permuté 2 | | .... |
|---|---|---|---|---|---|---|---|
| | réponse prédite | occurrences empirique | réponse prédite | occurrences empirique | réponse prédite | occurrences empirique | |
| A | **oui** | **15** | **oui** | **15** | non | 15 | |
| B | **oui** | **2** | non | 2 | **oui** | **2** | |
| C | non | 3 | non | 3 | non | 3 | |
| D | non | 4 | **oui** | **4** | non | 4 | |
| E | non | 2 | non | 2 | non | 2 | |
| F | non | 1 | non | 1 | non | 1 | |
| G | non | 0 | non | 0 | non | 0 | |
| H | non | 8 | non | 8 | **oui** | **8** | |
| I | non | 3 | non | 3 | non | 3 | |
| J | non | 2 | non | 2 | non | 2 | |
| | 17 occurrences capturées | | 19 occurrences capturées<br>19>=17 : oui | | 10 occurrences capturées<br>10>=17 : non | | .... |

Figure 29. Illustration d'un test de permutation sur la base d'un exemple fictif dans lequel 10 réponses sont à prédire. Chaque ligne correspond à une réponse possible. Des modèles permutés sont construits, produisant le même nombre de prédictions que le modèle originel. Nous regardons alors dans quelle proportion de cas un modèle ainsi généré obtient des résultats au moins aussi bons que le modèle originel.

8.3.2.5 Stratification du test par permutations

Cette analyse peut être faite à plusieurs niveaux. Le niveau « catégorie par catégorie » est le plus simple, le test de permutation est alors effectué sur chacune des catégories du modèle, mais cette approche soulève le risque de faire de multiples tests. Il est possible de tester le modèle de manière globale en effectuant une permutation des prédictions sur l'ensemble des prédictions au travers des différentes catégories. Cependant, pour éviter le biais soulevé plus tôt, cette permutation doit être contrainte de telle sorte que le **même nombre de prédictions soit fait dans chaque catégorie**, pour ne pas avantager injustement le modèle. En d'autres termes, l'espace de permutation correspond au produit cartésien des espaces de permutation dans chacune des strates (Pesarin & Salmaso, 2010, p. 38). Donc un modèle « **permuté par strates** » fait exactement le même nombre de prédictions pour chacune des catégories que le modèle original. Le principe reste ensuite le même : pour chaque modèle « permuté par strates » le nombre de réponses capturées est calculé et comparé au nombre de réponses capturées par notre modèle. La p-value correspond à la proportion de modèles « permutés par strates » qui capturent au moins autant de réponses que le modèle original.

#### 8.3.2.6 Tailles d'effet

Reporter seulement la p-value est une mauvaise pratique dans la recherche en psychologie cognitive, dans la mesure où – à taille d'effet constant – elle diminue avec la taille de l'échantillon. Nous verrons par ailleurs que la taille des données varie beaucoup selon les catégories étudiées, ce qui rend le problème plus important. Ainsi, APA 5e édition, insiste sur le report de tailles d'effet et d'intervalles de confiance (Fidler, 2002). Pour cette raison, nous avons choisi de construire une mesure intuitive de la taille de l'effet appelée « Force Relative » que nous représentons par l'acronyme **FR**. Elle mesure le rapport du nombre d'occurrences capturées par une formule prédite sur ce même nombre, mais pour une formule non prédite. Par exemple, une Force Relative de 3 indique que les formules prédites capturent trois fois plus d'occurrences que les formules non prédites. Un intervalle de confiance autour de cette valeur peut être estimé par une procédure dite de « bootstrap ». Sans rentrer dans les détails, cette procédure se rapproche du test de permutation, car il y a reconstruction d'un échantillon par le biais d'un tirage aléatoire (ici, encore 100 000). À la différence de ce dernier, le nouvel échantillon est tiré « *avec remise* ». Cette procédure est adaptée au calcul d'intervalle de confiance dans des cas non paramétriques (Efron & Tibshirani, 1986, p. 70).

#### 8.3.2.7 Méthode comparative

Une première possibilité pour comparer des modèles est simplement de comparer les valeurs dont nous avons parlé précédemment. Le premier problème avec cette méthode est qu'elle ne permet pas de déterminer si un modèle est significativement meilleur statistiquement parlant. Cette approche est par ailleurs problématique dans la mesure où il est fréquent que les modèles partagent un certain nombre de prédictions. Dans le cas particulier des modèles emboîtés, comme le modèle à une réinterprétation et le modèle à deux réinterprétations, un extrême est atteint, car toutes les prédictions de l'un sont des prédictions de l'autre. Ce n'est donc pas une technique précise pour comparer les différents modèles. Par ailleurs, il est réaliste de penser que les prédictions du modèle étendu sont moins puissantes que le modèle restreint, ce qui rend la comparaison directe caduque, car ses p-values et sa Force Relative peuvent avoir tendance à décroître.

De ce fait, il est désirable de centrer nos analyses sur les prédictions marginales qui sont réalisées. Le but n'est donc pas de savoir si le modèle 2 est meilleur que le modèle 1 du point de vue des métriques, mais plutôt de savoir si les prédictions qui distinguent les modèles entre eux ont une valeur statistique notable. Pour les étudier, il suffit de créer

un nouvel ensemble de données dans lequel ont été supprimées toutes les prédictions du modèle 2 qui sont partagées avec le modèle 1. Les réponses des élèves qui étaient appareillées à ces prédictions sont aussi supprimées. Il reste donc un ensemble de réponses d'élèves, que nous nommons les « inexpliqués du modèle 1 » et un ensemble de prédictions, que nous nommons « les prédictions marginales du modèle 2 ». Ce sont ces deux ensembles qui seront considérés. Ainsi, en utilisant la même méthode basée sur les tests de permutation, nous pouvons étudier si les prédictions marginales du modèle 2 sont meilleures que le hasard. Nous nommons par la suite cette méthode « la méthode soustractive »

## 8.4 Résultats

### 8.4.1 Répartition des erreurs

Comme nous l'avons indiqué dans un paragraphe précédent, les erreurs suivent des répartitions radicalement différentes selon leur catégorie d'appartenance. 1769 des 1904 erreurs se répartissent dans les catégories 1C (« un calcul ») et 2C (« deux calculs »).

Tableau 12. Répartition des erreurs selon les catégories de réponses. Les contraintes C1 (ne pas réutiliser de nombre) et C2 (utiliser tous les nombres) sont représentées.

| Nom de la catégorie | C1 respectée | C2 respectée | Nombre d'occurrences |
|---|---|---|---|
| 1C | oui | non | 1031 |
| 2C | oui | oui | 738 |
| 2C1R | non | non | 74 |
| 3C | non | oui | 61 |

Dans la réflexion que nous avons construite dans la partie théorique, les contraintes C1 et C2 sont préférentiellement respectées. Ce qui conduirait à un nombre important de réponses dans la catégorie 2C (qui ne peuvent alors être que des erreurs du fait de la particularité des problèmes rencontrés). Ainsi, nous nous attendions à un nombre plus important de réponses dans la catégorie 2C que dans toutes les autres catégories. Conformément à nos attentes, cette catégorie d'erreur est très populaire (738 occurrences), même si on note que la catégorie 1C l'est encore plus (1031 occurrences). Au contraire, la catégorie 2C1R et 3C ne représentent qu'une très faible partie des réponses. Cette contrainte d'utilisation des nombres semble donc être responsable d'une pression qui pousse le sujet à réinterpréter des phrases du problème plutôt que relâcher

ces contraintes. Compte tenu de la fréquence des observations, il est d'ores et déjà possible de suggérer que la contrainte de non-réutilisation de nombre a un poids plus important que celle d'utilisation de tous les nombres qui ne semble pas jouer un rôle important. En d'autres termes, ils acceptent l'idée que tous les nombres ne sont pas forcément indispensables à la résolution.

Cette répartition très inégale des erreurs appelle deux commentaires : la stratification proposée doit absolument être relativisée vis-à-vis du nombre d'occurrences dans ces catégories. En effet, il est « moins gênant » d'obtenir des résultats discutables dans la catégorie 3C et 2C1R, car elles représentent moins de 10 % des erreurs observées. Pour cette raison même de sous-effectif, le manque de puissance statistique pourra aussi être la cause de résultats non significatifs sans que l'effet analysé y soit moins présent.

## 8.4.2 Modèles de réinterprétations

L'ensemble des résultats portant sur les modèles de réinterprétations est résumé dans le Tableau 13.

### 8.4.2.1 Modèle des réinterprétations « de base »

Le premier modèle testé produit une seule réinterprétation pouvant avoir lieu n'importe quand au cours de la résolution. Le modèle effectue 211 prédictions sur 5893 possibles et capture 1071 observations sur un total de 1904. Comme précisé précédemment, il faut évaluer ce résultat en prenant en compte que les différentes catégories de formules ne se valent pas. Le test de permutation stratifié calculant la probabilité d'un modèle de tirer ses prédictions au hasard tout en conservant le même nombre de prédictions par catégories de formules permet d'attester de la significativité du pouvoir prédictif du modèle. En effet sur les 100 000 permutations testées, aucune n'a obtenu un meilleur score que le modèle de réinterprétation.

En effectuant le même type de test de permutation catégorie par catégorie, l'essentiel du succès du modèle provient de ses capacités prédictives sur la catégorie « un calcul » (1C). En effet les prédictions capturent 6 fois plus d'observations que les non-prédictions. La p-value du test de permutation est tout aussi forte que pour le test global. Les observations de la catégorie « deux calculs, sans réutilisation de quantité » (2C) sont aussi bien expliquées par le modèle avec une p-value équivalente, mais une Force Relative plus petite, estimée proche de 3.

Tableau 13. Évaluation des modèles de réinterprétations. IC est l'intervalle de confiance de la Force Relative établie par une procédure de bootstrap à 100 000 itérations.[37]

**MRD : Modèle de Réinterprétation direct**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 174/5893 | 1063/1904 | Sans objet | 0 |
| 1C | 37/160 | 677/1031 | 6.36(3.24-13.13) | 0 |
| 2C | 96/384 | 379/738 | 3.17(2.17-4.75) | 0 |
| 2C1R | 0/768 | 0/74 | Sans objet | 1 |
| 3C | 41/4581 | 7/61 | 14.35(0-41.25) | 0.00211 |

**MR1 : Modèle de Réinterprétation de base**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 211/5893 | 1071/1904 | Sans objet | 0 |
| 1C | 37/160 | 677/1031 | 6.36(3.22-13.16) | 0 |
| 2C | 96/384 | 379/738 | 3.17(2.17-4.74) | 0 |
| 2C1R | 29/768 | 8/74 | 3.09(0.78-7.04) | 0.06486 |
| 3C | 49/4581 | 7/61 | 11.99(0-34.68) | 0.00329 |

**MR1-MRD : Analyse des prédictions marginales**

| Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|
| 37/5719 | 8/841 | Sans objet | 0.06875 |
| 0/123 | 0/354 | Sans objet | - |
| 0/288 | 0/359 | Sans objet | - |
| 29/768 | 8/74 | 3.09(0.8-7) | 0.06684 |
| 8/4540 | 0/54 | 0 | 1 |

**MR2 : Modèle de Réinterprétation étendu**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 531/5893 | 1188/1904 | Sans objet | 0,00001 |
| 1C | 75/160 | 734/1031 | 2.8(1.34-6.78) | 0.00383 |
| 2C | 142/384 | 426/738 | 2.33(1.57-3.64) | 0,00003 |
| 2C1R | 97/768 | 16/74 | 1.91(0.83-3.89) | 0.11329 |
| 3C | 217/4581 | 12/61 | 4.93(1.06-11.62) | 0.00553 |

**MR2-MR1 : Analyse des prédictions marginales**

| Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|
| 320/5682 | 117/833 | Sans objet | 0.87464 |
| 38/123 | 57/354 | 0.43(0.21-1.02) | 0.84231 |
| 46/288 | 47/359 | 0.79(0.41-1.41) | 0.66677 |
| 68/739 | 8/66 | 1.36(0.43-3.03) | 0.29085 |
| 168/4532 | 5/54 | 2.65(0.42-6.6) | 0.11977 |

En ce qui concerne les deux dernières catégories, la significativité est plus difficile à atteindre du fait d'un manque de puissance (135 occurrences réparties sur ces deux catégories contre 1769 sur les deux suivantes). Il est ainsi difficile de conclure sur l'efficacité du modèle sur la catégorie « *deux calculs, avec réutilisation de quantité* » (2C1R) avec une p-value de 0.064 et une Force Relative autour de 3. D'un point de vue bayésien, il pourrait être argumenté qu'il est difficile de croire, compte tenu des résultats sur les autres catégories, que le modèle n'est pas pertinent sur cette catégorie de réponse. En termes de puissance, il aurait fallu que le modèle soit très explicatif pour atteindre la significativité sur cette catégorie. C'est peut-être le cas pour la dernière catégorie « *3 calculs* » (3C) qui obtient une p-value significative (p-value=0.0045). La

[37] La p-value nulle (p-value = 0) est un abus de langage. Il signifie ici que sur les 100 000 permutations pas une n'a obtenu un nombre de captures supérieur au modèle

Force Relative estimée est donc importante (évaluée à 12) et la significativité est atteinte malgré le faible nombre d'observations à expliquer (61).

#### 8.4.2.2 Comparaison avec les modèles alternatifs de réinterprétation

##### *8.4.2.2.1 Modèle à deux réinterprétations*

Deux versions du modèle de réinterprétations ont été testées : la première fait appel à une deuxième réinterprétation. Les performances globales de ce deuxième modèle sont très bonnes, cependant elles le sont moins sur chacune des catégories que le modèle de base. En effet, dans ce premier modèle, la significativité des prédictions est toujours atteinte pour toutes les catégories sauf pour (2C1R). Nous avons argumenté plus tôt les limites de la comparaison des modèles par la mise en correspondance des métriques calculées. Par la méthode soustractive que nous avons détaillée, nous comparons le modèle à une réinterprétation avec le modèle à deux réinterprétations.

Les résultats vont très clairement à l'encontre de ces nouvelles prédictions. En effet, dès l'évaluation globale du modèle par un test de permutation stratifié, la p-value obtenue est de 0.87 indiquant donc que le hasard était meilleur dans plus de 8 cas sur 10 que les prédictions originales produites par le modèle avec deux réinterprétations. Les analyses catégorie par catégorie ne permettent pas vraiment de nuancer les mauvaises performances du modèle. Si la dernière catégorie peut laisser un doute (p-value = 0.12, FR = 2.65), le hasard aurait pu facilement produire ce résultat, une étude avec plus de puissance aiderait à élucider si oui ou non, les doubles réinterprétations peuvent expliquer partiellement les formules de ces catégories.

##### *8.4.2.2.2 Modèle aux réinterprétations directes*

La deuxième version du modèle, *a contrario* de la première, produit une seule réinterprétation et seulement **dès le début de la résolution**. Ses prédictions sont plus puissantes dans la mesure où elles ne représentent pas les cas qui ne peuvent être obtenus qu'avec une réinterprétation tardive. Ces solutions correspondent à des chemins plus rares dans l'ensemble des résolutions possibles. Nous obtenons en effet de très bonnes p-values, tant au niveau du test global (p-value < 1.10^-5) que les différentes catégories (1C : p-value < 0.00001; 2C : p-value < 0.00001; 3C : p-value=0,0024). Lorsque la comparaison est réalisée avec la méthode soustractive, les prédictions marginales du modèle « de base » ne passent pas la significativité (cf. Tableau 13). Mais il convient de noter que les modèles « direct » et « normal » diffèrent assez peu et qu'il en résulte un manque de puissance notable. En effet, seulement 37 nouvelles prédictions

sont produites par le modèle « de base ». Cependant ce modèle « direct » ne peut pas produire de réponses dans la catégorie 2C1R[38]. Les statistiques pour cette catégorie restent encourageantes, même si la p-value n'est pas significative (p-value=0.0648). En ce qui concerne la catégorie 3C, seules 8 prédictions sont présentes, et aucune n'est réalisée, ce qui implique une p-value de 1 et une Force Relative de 0. Ici aussi, le manque de puissance statistique ne permet pas vraiment de tirer de conclusion solide : à titre de comparaison, le modèle normal tire 49 prédictions sur cette catégorie, pour capturer seulement 7 observations.

En conclusion, il existe un manque important de puissance, car les formules explicables seulement par des réinterprétations tardives sont au nombre de 37, soit moins d'un cinquième des prédictions du modèle « de base ». Nous choisissons de conserver le modèle de base comme modèle de référence pour simuler les réinterprétations dans la résolution, car si statistiquement la différence est faible, l'idée de pouvoir faire des prédictions dans la catégorie 2C1R est une propriété que nous souhaitons conserver.

### 8.4.3 Modèles des mots-clefs

[38] Cet aspect de la catégorie 2C1R vis-à-vis des modèles de réinterprétation est assez subtil. En effet, chaque quantité du problème appartenant à un unique schéma, il faut qu'une réinterprétation change son rôle **après** un calcul qui l'utilise correctement pour pouvoir produire une formule de cette catégorie.

Tableau 14. Évaluation et comparaison du modèle des mots-clefs à liste restreinte et à liste étendue.

**MMR : Modèle de mots-clefs liste restreinte**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 738/5893 | 1090/1904 | Sans objet | 0 |
| 1C | 47/160 | 868/1031 | 12.8(7.89-19.43) | 0 |
| 2C | 138/384 | 210/738 | 0.71(0.46-1.07) | 0.94035 |
| 2C1R | 96/768 | 7/74 | 0.73(0.2-1.63) | 0.68243 |
| 3C | 457/4581 | 5/61 | 0.81(0.17-1.83) | 0.62143 |

**MME : Modèle de mots-clefs liste etendue**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 1042/5893 | 639/1904 | Sans objet | 0.88116 |
| 1C | 76/160 | 386/1031 | 0.66(0.31-1.38) | 0.85801 |
| 2C | 128/384 | 245/738 | 0.99(0.62-1.51) | 0.509 |
| 2C1R | 160/768 | 4/74 | 0.22(0-0.58) | 0.99678 |
| 3C | 678/4581 | 4/61 | 0.4(0.08-0.96) | 0.92494 |

**MME-MMR : Analyse des prédictions marginales**

| Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|
| 657/5155 | 164/814 | Sans objet | 0.96246 |
| 43/113 | 33/163 | 0.41(0.23-0.67) | 0.99951 |
| 65/246 | 127/528 | 0.88(0.46-1.52) | 0.66249 |
| 96/672 | 3/67 | 0.28(0-0.86) | 0.96536 |
| 453/4124 | 1/56 | 0.15(0-0.55) | 0.98268 |

Les performances du modèle des mots-clefs restreint ont une variabilité notable. Le test de permutation global assure la pertinence du modèle avec une p-value < 1.10^-5. Cependant, contrairement au modèle des réinterprétations, ses bonnes performances ne sont concentrées que sur la catégorie « un seul calcul » (1C). Sur cette catégorie, les prédictions sont très bonnes avec une p-value< 1.10^-5 et une Force Relative de 12.8. Au contraire, pour les autres catégories, elles sont plutôt mauvaises, avec des p-values de 0.94, 0.68, 0.62 respectivement pour les catégories 2C, 2C1R, 3C et des forces relatives de 0.71, 0.73 et 0.81. Un manque de puissance ne peut être imputé, car le nombre de prédictions pour ces 3 catégories est élevé : 138, 96, 457 versus 142, 97 et 217, respectivement, pour le modèle de réinterprétation.

Les performances globales pour le modèle par mots-clefs étendu ne sont pas bonnes. En effet, d'une part, les performances sur ces trois catégories ne sont pas meilleures et le test de permutation stratifié ne donne pas de résultat significatif (p-value = 0.883). De plus, les prédictions sur la catégorie 1C, à la différence du premier modèle par mots-clefs, ne sont pas non plus concluantes. Par construction, les prédictions du modèle par mots-clefs étendu ne contiennent pas toujours l'ensemble des prédictions du modèle par mots-clefs à liste restreinte. C'est ce qui explique la médiocrité inattendue des performances du modèle étendu. De ce fait, il est important de pouvoir identifier si les nouvelles prédictions du modèle étendu ont une quelconque valeur, ce que fait notre méthode soustractive. Deux cas sont possibles :

- Si les nouvelles prédictions sont pertinentes, alors il faut réfléchir à une construction de ce modèle qui n'enlève pas certaines prédictions du modèle restreint.
- Si elles ne le sont pas, il n'est pas utile de chercher à le reconstruire intégralement.

L'analyse des prédictions marginales de ce modèle permet de trancher la question vers la deuxième option. Des tests non significatifs sont trouvés, que ce soit pour le test global (p-value = 0.96) ou pour chacune des différentes catégories (p-values entre 0.99 et 0.66).

### 8.4.4 Comparaison du Modèle des réinterprétations et du modèle des mots-clefs.

Dans les parties précédentes, nous avons comparé trois versions du modèle des réinterprétations d'une part, et deux versions du modèle des mots-clefs d'autre part. Nous présentons maintenant la comparaison du modèle des mots-clefs restreint au modèle des réinterprétations. Pour les deux modèles, les performances globales sur la catégorie (1C) étaient très bonnes. Pour les autres catégories, le modèle des mots-clefs n'est pas explicatif alors que le modèle des réinterprétations est assez prédictif. Si nous concentrons notre attention sur la première catégorie (1C), étant donné que ces deux modèles partagent beaucoup de prédictions sur cette catégorie, il est important d'établir si leurs prédictions marginales sont pertinentes ou si l'un est radicalement plus explicatif que l'autre.

Notre méthode comparative va nous permettre d'éclairer cette question.

Tableau 15. Comparaison deux à deux du modèle de réinterprétation « de base » et du modèle de mots-clefs à liste restreinte.

**MR-MMR : Analyse marginale**
**Modèle des réinterpretations Vs Inexpliqués du modèle des mots-clefs**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 139/5155 | 322/814 | 23.62 | 0 |
| 1C | 14/113 | 43/163 | 2.53(1.28-4.22) | 0.0013 |
| 2C | 57/246 | 266/528 | 3.37(2.15-5.54) | 0 |
| 2C1R | 20/672 | 6/67 | 3.21(0.44-8.3) | 0.07701 |
| 3C | 48/4124 | 7/56 | 12.13(0-35.72) | 0.00347 |

**MMR-MR : Analyse marginale**
**Modèle des mots clefs Vs Inexpliqués du modèle des réinterpretations**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 666/5682 | 341/833 | 5.22 | 0.00653 |
| 1C | 24/123 | 234/354 | 8.04(2.61-15.86) | 0.00003 |
| 2C | 99/288 | 97/359 | 0.71(0.35-1.33) | 0.83336 |
| 2C1R | 87/739 | 5/66 | 0.61(0.1-1.55) | 0.74885 |
| 3C | 456/4532 | 5/54 | 0.91(0.19-2.1) | 0.54961 |

L'analyse réalisée montre que les 14 prédictions marginales du modèle des réinterprétations sont meilleures que le hasard (p-value = 0.0019, FR = 2.53). C'est aussi le cas pour le modèle des mots-clefs, avec des résultats encore plus solides (p-value < 0.00001, FR = 8). Les deux modèles ont donc un pouvoir prédictif qui leur est propre. Il est aussi possible d'effectuer des comparaisons sur les autres catégories, les résultats sont évidemment sans surprise, car le modèle des mots-clefs est hors jeu sur ces trois autres catégories et le modèle des réinterprétations a de bonnes performances sur celles-ci.

### 8.4.5 Comment expliquer les faibles performances du modèle des mots-clefs étendu ? Analyses Post-Hoc

Dans un test de permutation, il est possible d'inverser le raisonnement. La proportion des cas pour lesquels le modèle originel est moins bon qu'un modèle permuté peut être

calculée pour tester s'il est **moins bon que le hasard**. Si un tel test devait être fait pour les prédictions marginales du modèle étendu, alors il aurait atteint la significativité dans la catégorie 1C avec une p-value inférieure[39] à 0,00049.

Ce résultat est surprenant dans la mesure où le modèle par mots-clefs à liste étendue est construit sur des bases identiques au modèle à liste restreinte qui, lui, obtient de très bons résultats. Nous avons donc inspecté « manuellement » ce que ces nouvelles prédictions avaient de spéciales et nous avons formé et testé une hypothèse post-hoc pour la tester et contrôler qu'elle ne remette pas en cause les autres modèles.

#### 8.4.5.1 Présence d'une soustraction

En sélectionnant les formules dans la catégorie 1C sur lesquelles le modèle restreint ne fait pas de prédiction, il est possible de remarquer une différence importante entre les formules prédites par le modèle étendu et celles qui ne sont pas prédites : la présence de soustraction. En effet, dans le modèle étendu, les prédictions comportant une soustraction sont rares alors que les non-prédictions comportent un grand nombre de soustractions. Pour la catégorie 1C, le modèle des mots-clefs étendu rajoute **33 additions et aucune soustraction dans la liste des prédictions**. Or, ces additions sont clairement sous-représentées dans les réponses des élèves.

Deux questions se posent alors : (1) pourquoi le modèle étendu ne propose pas de soustraction ? (2) Pourquoi les soustractions sont-elles plus populaires ? En reprenant un problème comme celui qui suit :

> *En janvier, 6 enfants se sont inscrits à la chorale.* ***Après*** *janvier, il y a 13 enfants à la chorale. Avant janvier, il y avait autant d'enfants inscrits au football qu'à la chorale. En janvier, il y a eu de nouvelles inscriptions au football.* ***Après*** *janvier, il y a 2 enfants de moins au football qu'à la chorale. Combien d'enfants se sont inscrits au football en janvier ?*

[39] « strictement inférieur », car il faut compter les cas où le modèle permuté fait le même nombre de captures que le modèle testé. Donc la p-value correspondante est plus petite que 1-p avec p, la p-value du test de permutation sur la catégorie 1C (p-value =0,99951).

Il apparait que les deuxième et troisième nombres sont indicés par une addition expliquant de nouvelles formules, car le mot-clef « après » fait partie de la liste étendue de mots-clefs. C'est le cas pour la plupart des problèmes. Ces nouveaux mots-clefs vont donc avoir pour effet de bord de compter moins de soustraction dans les prédictions du modèle.

Notre hypothèse sur l'importance de la soustraction est que le mot-clef mis en valeur à la fin du problème a une force plus grande que les autres mots-clefs et pourrait, par exemple, impliquer des réponses comme T1-P1, même si le signe moins ne concernait que la quantité « d ». S'il est difficile d'élucider ce problème avec nos données actuelles, nous pouvons quand même étudier si « faire une soustraction » est effectivement une contrainte importante dans la résolution.

#### 8.4.5.2 Tester l'hypothèse par la construction d'un nouveau modèle minimaliste

Pour tester l'hypothèse ad hoc selon laquelle les élèves tendent à respecter la contrainte d'effectuer une soustraction, nous avons, de manière analogue aux modèles précédents, construit et testé la pertinence des formules qui respectent cette contrainte. Ce n'est pas un modèle à proprement parler, car nous n'avons pas de théorie de construction de solution, mais plutôt le test d'une contrainte. Cette contrainte se traduit simplement par « la résolution passe par au moins une soustraction ». Ces formules sont, de manière étonnante, très prédictives. Nous obtenons des p-values significatives sur le test stratifié global, ainsi que sur toutes les catégories (cf. Tableau 16). Ce résultat laisse penser que cette contrainte est présente. Évidemment le nombre de prédictions est très grand, mais largement compensé par le taux important de captures.

Tableau 16. Analyse des performances du modèle prédisant la présence de soustraction et comparaison aux modèles des mots-clefs et des réinterprétations.

**MS : Modele predisant la Soustraction**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 4933/5893 | 1716/1904 | 1.78 | 0 |
| 1C | 64/160 | 915/1031 | 11.83(7.14-18.8 | 0 |
| 2C | 288/384 | 669/738 | 3.23(2.1-5.48) | 0 |
| 2C1R | 576/768 | 71/74 | 7.89(3.08-Inf) | 0.00048 |
| 3C | 4005/4581 | 61/61 | Sans Objet | 0.00444 |

**MMR-MS : Analyse marginale**
**Modèle des mots clefs Vs Inexpliqués du modèle de soustraction**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 114/960 | 61/188 | 3.56 | 0.00089 |
| 1C | 17/96 | 43/116 | 2.74(1.42-4.72) | 0.00113 |
| 2C | 15/96 | 18/69 | 1.91(0.52-4.42) | 0.11114 |
| 2C1R | 22/192 | 0/3 | 0(-) | 1 |
| 3C | 60/576 | 0/0 | Sans Objet | 1 |

**MR-MS : Analyse marginale**
**Modèle des réinterprétations Vs Inexpliqués du modèle de soustraction**

| Espace | Npred/Nposs | Ncapt/Ntot | FR(IC) | p-value |
|---|---|---|---|---|
| Global | 16/960 | 48/188 | 20.23 | 0.00006 |
| 1C | 16/96 | 48/116 | 3.53(1.9-5.91) | 0.00004 |
| 2C | 0/96 | 0/69 | Sans Objet | 1 |
| 2C1R | 0/192 | 0/3 | Sans Objet | 1 |
| 3C | 0/576 | 0/0 | Sans Objet | 1 |

8.4.5.3 Quelles conséquences sur les autres modèles ?

Au vu de cette observation, il est incontournable de réétudier les modèles précédents. Nous devons nous assurer que nous ne sommes pas dans la situation où les modèles précédents sont prédictifs du fait qu'ils prédisent une soustraction. Toujours par la méthode soustractive, des analyses ont été faites pour s'assurer de la pertinence des prédictions sans soustraction (cf. Tableau 16). Pour les deux modèles (mots-clefs et réinterprétations), la valeur des modèles est effectivement conservée tant d'un point de vue global, que sur la catégorie 1C. Pour les catégories suivantes, le modèle des mots-clefs reste pertinent, mais le modèle des réinterprétations ne peut pas être jugé dans la mesure où il ne prédit pas non plus les formules consistant à tout additionner. La raison pour la non-prédiction de formules ne comportant que des additions vient du fait que la structure mathématique des problèmes n'a pas plusieurs niveaux d'imbrication.

Autrement dit, même lorsqu'il y a des réinterprétations de quantités, il n'y a pas de raison d'enchaîner plusieurs additions pour arriver au résultat. Quand bien même le modèle ferait ce type de prédiction, il est absolument impossible de juger de la pertinence, car il n'y a aucune observation qui rentre dans la catégorie 3C, et seulement 3 dans la catégorie 2C1R.

### 8.4.6 Synthèse des résultats

La Figure 30 décrit un résumé des résultats de cette section. Outre les p-values globales satisfaisantes, nous pouvons noter un bon niveau de compatibilité entre les modèles différents. Les mots-clefs, la nécessité d'une soustraction (que nous avons imputée au mot-clef de la fin) et les réinterprétations semblent tous trois jouer un rôle important dans la résolution.

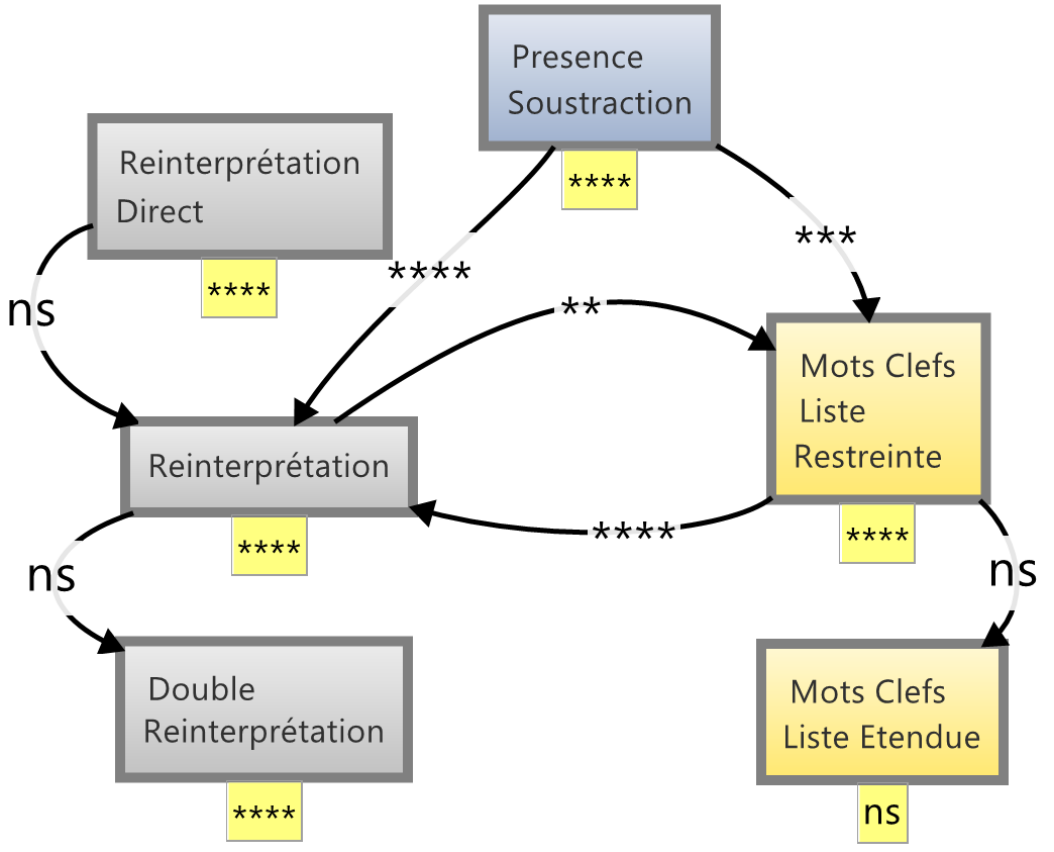



Figure 30. Présentation synthétique des tests de permutation stratifiés globaux sur les modèles et sur les comparaisons de modèle. Une flèche pleine indique une soustraction du modèle à sa tête par le modèle à sa queue. Légende : ns = non significatif; ** = $p<0.01$ ; *** = $p<0.001$ ; **** = $p<0.0001$

## 8.5 Discussion

### 8.5.1 Évaluation des modèles

Nous avons montré que le modèle de réinterprétation de base a un bon pouvoir prédictif pour expliquer les observations présentes dans les différentes catégories de réponses. Le modèle des mots-clefs, quant à lui, a un fort pouvoir explicatif des réponses appartenant à la première catégorie, mais se révèle inadéquat pour les trois autres catégories. Par ailleurs, nous avons testé des versions alternatives de ces deux différents modèles. Notre méthode soustractive nous a permis de déterminer si les prédictions originales d'un modèle étendu sont meilleures que le hasard, ou au contraire, s'il n'apparait explicatif que parce qu'il partage des prédictions avec son homologue restreint. Nous avons mis en évidence que, pour que le modèle des mots-clefs fasse des prédictions pertinentes, il fallait qu'il soit constitué d'une liste relativement restreinte de mots-clefs candidats. En ce qui concerne le modèle des réinterprétations, ajouter la possibilité d'une deuxième réinterprétation ne permettait pas d'expliquer plus efficacement les observations. Nous avons aussi pu tester un modèle plus restreint que le modèle de base produisant des réinterprétations seulement en début de résolution. Nous avons noté que les prédictions entre ce modèle et les modèles de base diffèrent peu (37 prédictions différentes seulement). Il est ainsi difficile de juger de façon statistique si les prédictions concernant les formules passant par une interprétation tardive sont suffisamment prédictives. Cependant, elles portent intégralement sur une catégorie de réponses qui, sans elles, resteraient inexpliquées.

À propos du lien entre les interprétations et les contraintes d'utilisation des nombres, nous nous attendions à une relation plus claire : plus de réponses appartenant à la catégorie ne relâchant aucune contrainte d'utilisation des nombres (catégorie 2C), et présence d'un phénomène fort de réinterprétation. Or il est apparu que cette catégorie est moins populaire que la catégorie 1C. D'autre part, la Force Relative des interprétations sur cette catégorie semble être plus élevée que celle sur la catégorie « deux calculs ». Cependant, le modèle des mots-clefs a fourni d'excellents résultats inattendus. Nos études comparatives ont montré que ni le modèle des réinterprétations ni celui des mots-clefs ne peuvent être vus comme dominés par l'autre sur la catégorie « un calcul » puisque les prédictions marginales effectuées par l'un ou l'autre modèle restent largement meilleures que le hasard.

### 8.5.2 Limites des travaux dues à la similarité des problèmes complexes.

La structure similaire des problèmes sur lesquels ont travaillé les enfants a constitué un atout, elle permet de se concentrer sur les variations dans la formulation des relations entre les quantités. Elle cause cependant une limite claire sur la généralisation de nos résultats. En effet, elle peut être responsable de tendances générales difficiles à contrôler. De ce point de vue, notre démarche présente donc certaines limites. Le fait que les nouvelles prédictions du modèle des mots-clefs étendu soient significativement moins bonnes que le hasard nous a conduits à établir et à tester une hypothèse ad hoc en testant un modèle composé uniquement de la contrainte consistant à utiliser au moins une soustraction au cours de la résolution. En testant cette hypothèse de manière indépendante, elle semble effectivement validée et détenant un fort pouvoir explicatif. Une interprétation possible est de considérer cette soustraction comme une contrainte à part entière dans la résolution du problème, autrement dit, les mots-clefs ne sont pas seulement un mode de résolution possible, mais peuvent aussi être interprétés comme des contraintes que l'élève s'efforce de respecter. Ainsi le mot-clef se situant à la fin du problème imposerait une contrainte sur la résolution. Une autre hypothèse serait de remettre en cause le choix de modélisation consistant à associer les mots-clefs aux nombres de la même phrase. Seul un nouveau travail de modélisation permettrait d'évaluer la pertinence de ces deux hypothèses.

Par exemple, dans les problèmes fournis, une comparaison du type « d de moins » est toujours utilisée à la fin. La présence d'une comparaison de type « d de plus » dans certains énoncés aurait permis de tester avec plus de précision l'hypothèse selon laquelle le mot-clef le plus mis en valeur a une portée globale sur la résolution.

En termes de perspectives, il serait intéressant d'utiliser des données plus riches pour accompagner cette méthode d'analyse. Par exemple, il serait intéressant de demander aux élèves leur niveau de confiance pour chacune de leurs réponses. Nous pourrions construire l'hypothèse selon laquelle les réponses passant par une réinterprétation sont plus marquées de confiance que les réponses utilisant les mots-clefs. D'autres mesures sont aussi possibles comme le temps de réponse, en faisant l'hypothèse que les réinterprétations ont lieu dans des situations d'impasse détectées par des pauses dans la résolution. Les mesures de temps de réponse permettraient aussi de savoir si l'utilisation des mots-clefs est généralement un pilotage par défaut (au sens d'une stratégie utilisée directement) ou si elle est plutôt la résultante d'un relâchement de contraintes faisant suite aux réinterprétations.

Enfin au niveau du matériel, il serait intéressant de construire des problèmes permettant d'obtenir des prédictions les plus éloignées possible entre les modèles de réinterprétation et des mots-clefs pour augmenter la puissance de la comparaison. Il serait ainsi plus facile de discerner si une réponse provient d'un travail sur les mots-clefs ou d'une réinterprétation des problèmes. Toujours au niveau du matériel, il serait possible d'utiliser des problèmes qui poussent un peu plus à la réinterprétation. Nous avons vu que la contrainte d'utiliser tous les nombres est finalement plus souvent relâchée que celle de ne pas réutiliser de nombre. De ce fait, les problèmes à deux nombres demandant trois calculs[40] devraient plus conduire à la réinterprétation de certaines phrases du problème. Il serait intéressant de tester cette hypothèse. Reproduire ces travaux avec des problèmes plus variés est donc une nécessité pour pouvoir être en mesure de généraliser ces résultats. Tester un modèle contre le hasard a ses limites dans la mesure où il peut faire de bonnes prédictions pour de mauvaises raisons, ainsi il est capital de pouvoir mettre à l'épreuve non pas de nouvelles données, mais de nouveaux modèles, donc de continuer à employer la méthode comparative utilisée dans cette partie.

D'autres modèles de la résolution peuvent être construits. Il est important, si nous voulons chercher à valider proprement[41] le modèle des réinterprétations, de le mettre en compétition avec des modèles qui partagent un recouvrement naturel fort avec ses prédictions. En effet, nous soulignons que les réinterprétations du modèle permettent des opérations entre des quantités qui partagent un contexte sémantique commun. Ainsi il est fréquent, dans les problèmes fournis, que le premier nombre et le dernier nombre appartiennent à des contextes sémantiques très différents, donc les réinterprétations prédisent rarement la possibilité d'une opération entre ces quantités. Cela conduit à une abduction importante du champ des formules attendues. Cette suppression pourrait être partagée par des modèles construits sur des bases différentes que le modèle des réinterprétations, mais qui seraient touchés de la même manière par cette réduction.

---

[40] Par exemple « Marie a 5 billes, Jean en a 6 de plus. Combien ont-ils de billes ensemble ».

[41] Ou bien chercher à le falsifier, si nous nous plaçons dans une démarche scientifique au sens de Popper.

### 8.5.3 Existence probable de meilleurs modèles

Bien que nos choix aient été guidés par l’analyse de la littérature sur le sujet, il est possible que certaines réinterprétations possibles aient été oubliées ou que certaines soient obsolètes. La même remarque peut être établie pour le modèle des mots-clefs. De meilleures prédictions peuvent être attendues une fois ces modèles affinés. Enfin, nous rappelons que le modèle génère des réponses de manière combinatoire. Il serait intéressant de faire intervenir des contraintes supplémentaires guidant la résolution, pour simuler moins de réponses. Tout comme les travaux de modélisation dans le premier chapitre, des sets de contraintes propres au sujet peuvent intervenir.

# 9 MODELISATION EPISTEMIQUE. DIAGNOSTIQUER L'APPRENANT.

Une attente vive en EIAH est de doter les technologies éducatives de la capacité de se centrer sur l'apprenant avec une capacité de diagnostic égale ou supérieure au tuteur humain. Dans les chapitres précédents, nous avons traité de la modélisation à un niveau comportemental puis à un niveau cognitif. Toutes deux participent à l'objectif de mieux diagnostiquer l'apprenant dans sa résolution de PAEV. Le premier chapitre des contributions représente la première étape dans la construction du diagnostic cognitif. Cette étape est le diagnostic comportemental, elle consiste à déterminer les actions prises par l'apprenant lors de sa résolution. La deuxième étape, nommée le diagnostic épistémique, constitue à attribuer des traits cognitifs à l'apprenant. Cette étape n'est accessible que si des modèles sont à dispositions pour mettre en lien ces traits et les réponses du sujet. Ainsi, nous avons consacré notre premier chapitre des contributions à la modélisation cognitive du sujet pour mieux connaître les processus en jeu lors de la résolution de PAEV. Nous avons cherché à modéliser les erreurs par le biais des réinterprétations et des mots-clefs. En plus de ses propres visées théoriques, ce travail a permis de créer une base de réponses possibles auxquelles nous pouvons associer une explication. Cependant, cette association est statistique et nous n'avons pas de certitudes

en ce qui concerne la réponse d'un élève particulier. Ainsi, est-il juste de se baser uniquement sur cette supposition pour, par exemple, produire un diagnostic ? Ce travail portait sur la mise en place d'un modèle cognitif au sens de la psychologie cognitive, mais dans le cadre des EIAH c'est l'établissement d'un modèle singulier de l'élève qui est visé dans le but de produire un diagnostic.

Dans le cadre d'un système de diagnostic, pour pouvoir parler de diagnostic, il faut avoir capturé une régularité dans les réponses de l'apprenant. Ainsi, certains environnements dédiés à l'algèbre comme Pépite (Delozanne et al., 2008; El-Kechaï et al., 2011) ou Aplusix (Hamid Chaachoua, Nicaud, Bronner, & Bouhineau, 2004) ont pour stratégie de laisser l'enfant répondre à plusieurs questions avant d'établir un diagnostic cognitif.

Si la question du diagnostic de l'individu est importante en EIAH, elle n'en est pas moins cruciale du point de vue de la recherche en psychologie. Est-il possible de parler de modélisation approfondie si les différences interindividuelles n'ont pas de place dans le modèle ? Dans le précédent chapitre, nous avons contribué à une meilleure connaissance de ces aspects de la résolution de problème, mais nous n'avons pas produit de modèle pouvant simuler un élève en particulier. Notre question de recherche est la suivante : les contraintes dans l'usage des nombres et la résolution par mot-clef ou par réinterprétation sont-elles une régularité de l'apprenant ? Ou au contraire, le niveau de compétence est-il un trait plus stable du comportement des apprenants ?

Dans la première moitié de cette partie, nous présentons un critère simple pour l'évaluation de diagnostic et de modèle de diagnostic qui permet de mesurer la qualité d'un diagnostic tout en pénalisant le nombre de paramètres libres. Un outil auteur, nommé **STAR** pour **S**imple **T**oolbox to **A**nalyse **R**easonning, est présenté. Il implémente le critère d'intérêt et permet la construction de modèles de diagnostic (Bruno Martin, Sander, Labat, & Richard, 2013). Il est particulièrement adapté pour décrire un modèle de diagnostic basé sur une analyse du comportement par contraintes que nous présentons par la suite. Dans la deuxième moitié, nous utilisons donc notre outil pour investiguer les différences interindividuelles dans l'arithmétique en classe élémentaire. Un module dans DIANE a été rajouté pour que les données puissent être analysées dans STAR. Nous avons construit un modèle simple permettant à la fois d'investiguer les comportements de résolution du point de vue des différences interindividuelles et de mettre en pratique STAR sur un problème de modélisation concret. Outre le souhait d'employer STAR sur un cas concret, cette expérimentation

vise à rechercher des différences interindividuelles sur la base de nos travaux théoriques précédents et sur la base méthodologique donnée par STAR.

## 9.1 Notre proposition

Nous avons consacré un passage au principe du Minimum Description Length dans le chapitre théorique consacré à l'évaluation des modèles. Ce principe est intéressant, car, contrairement aux critères comme l'AIC et BIC, il laisse des possibilités d'implémentations différentes. Nous avons noté dans cette partie, toutefois, qu'à notre connaissance aucune implémentation ne permet aux modèles cognitifs de préserver leur caractère non probabiliste. Nous proposons dans cette partie une implémentation de ce principe adaptée aux modèles déterministes.

### 9.1.1 Un MDL brut en deux parties

Notre proposition consiste à reprendre l'idée de complexité stochastique telle qu'elle est décrite par Rissanen (1986, p. 1080), mais en l'adaptant à la classe des modèles déterministes :

> *As a modification of the notion of algorithmic complexity, the stochastic complexity of a string of data, relative to a class of probabilistic models, is defined to be the fewest number of binary digits with which the data can be encoded by taking advantage of the selected models.*

Mesurer la qualité d'un modèle par un calcul de complexité stochastique est à la base du Minimum Description Length en deux parties. Cependant, comme indiqué plus haut, ces critères ne concernent que les modèles probabilistes et ne s'appliquent aux modèles déterministes que par le recours à une fonction d'erreur.

Nous proposons de reprendre cette définition, mais en évitant l'approche probabiliste, et en se basant sur l'idée qu'un modèle déterministe permet de décrire une version résumée des données auxquelles il est confronté. Nous décidons de généraliser l'idée du MDL brut en deux parties, c'est-à-dire compter le nombre de bits requis pour transmettre (1) les paramètres choisis par le modèle, (2) les données en tirant parti du modèle. Le programme MDLchunker (Robinet et al., 2011) représente une utilisation proche de celle que nous souhaitons mettre en place. Les auteurs font appel à la version en deux parties pour rendre compte de l'activité de chunking, c'est à dire stocker en mémoire des groupes d'éléments plutôt qu'un élément unique. Pour que cette activité soit pertinente, il faut qu'elle permette de simplifier la représentation à laquelle le sujet

est confronté. Une version brute du MDL permet d'établir à la fois le coup en bit de la création d'un chunk, et son apport, en bits, pour simplifier la représentation à laquelle le sujet est confronté.

## 9.1.2 Formulation simplifiée à partir d'un exemple

### 9.1.2.1 Description succincte des données

Le critère calculé est basé sur le principe du Minimum Description Length. Par abus de langage, notre score sera simplement décrit sous le nom « MDL ».

Avant de détailler le calcul du critère dans le cadre général, un exemple est présenté pour illustrer le principe de notre proposition. Il montre comment, en tirant profit du modèle de diagnostic, il est possible de décrire sous un format compressé les traces de l'apprenant.

Tableau 17. Données récupérées d'un tableau de Tatsuoka (2009, p. 29) illustrant la méthode RSM.

| Task Number* Example | Three Responses to Four Parallel Forms Within the 64-Item Test | | |
|---|---|---|---|
| | Student 1 | Student 2 | Student 3 |
| 6. 6 + 4 = +10 | 1111 (+10) | 1111 (+10) | 1111 |
| 15. –6 + 4 = –2 | 1111 (–2) | 0000 (–10) | 1111 |
| 3. 12 + –3 = +9 | 1111 (+9) | 0000 (+15) | 1111 |
| 5. –3 + 12 = +9 | 1011 (+9) | 0000 (+15) | 1111 |
| 10. –14 + –5 = –19 | 1111 (–19) | 1111 (–19) | 1111 |
| 11. 3 + –5 = –2 | 1111 (–2) | 0000 (–8) | 1111 |
| 14. –5 + –7 = –12 | 0000 (–14) | 1111 (–12) | 1111 |
| 7. 8 – 6 = +2 | 0000 (+14) | 0000 (+14) | 1111 |
| 8. –16 – –7 = –9 | 0000 (–23) | 0000 (–23) | 1111 |
| 16. 2 – 11 = –9 | 0000 (+13) | 0000 (+13) | 0111 |
| 13. –3 – + 12 = +9 | 0000 (+9) | 0000 (+15) | 1111 |
| 1. –6 – –8 = +2 | 0000 (–14) | 0000 (–14) | 1111 |
| 12. 9 – –7 = +16 | 0000 (+2) | 1111 (+16) | 0011 |
| 4. 1 – –10 = +11 | 0000 (–9) | 0000 (–11) | 1010 |
| 2. –7 – 9 = –16 | 0000 (+2) | 0000 (+16) | 1111 |
| 9. –12 –3 = –15 | 0000 (–9) | 1111 (–15) | 0111 |

Dans cet exemple, des règles erronées possibles sont imaginées à l'avance, et sont recherchées chez l'apprenant. Leurs réponses sur une batterie d'item sont données en Tableau 17. L'étudiant 1 est diagnostiqué avec la règle composite que nous nommons par la suite **règle 1**: *« Si deux nombres ont le même signe, alors les ajouter en valeur absolue et prendre ce signe »* et *« Si deux nombres ont des signes différents alors*

*soustraire le plus petit au plus grand et prendre le signe du nombre qui a la plus grande valeur absolue* ». L'étudiant 2 est diagnostiqué avec la règle 2 « *ajouter les deux nombres et prendre le signe du nombre qui a la plus grande valeur absolue* ». L'étudiant 3 est diagnostiqué comme *suivant le modèle expert.* Chaque étudiant est passé 4 fois sur chaque type de problème. Nous notons 1 si réussite et 0 si échec. Le nombre donné par l'élève est entre parenthèses. Dans le modèle de Tatsuoka, cette valeur n'est pas réutilisée (autrement que par son caractère juste/faux). Les règles se différencient par leurs variations dans la prédiction de réussite selon les différents problèmes.

9.1.2.2 Données non compressées

La complexité des données peut être réduite si les règles prévues sont respectées. Tout d'abord quelle serait la taille du message si l'ensemble des réponses du sujet 1 devait être transmis en langage binaire ? La Figure 31 propose une description simple :

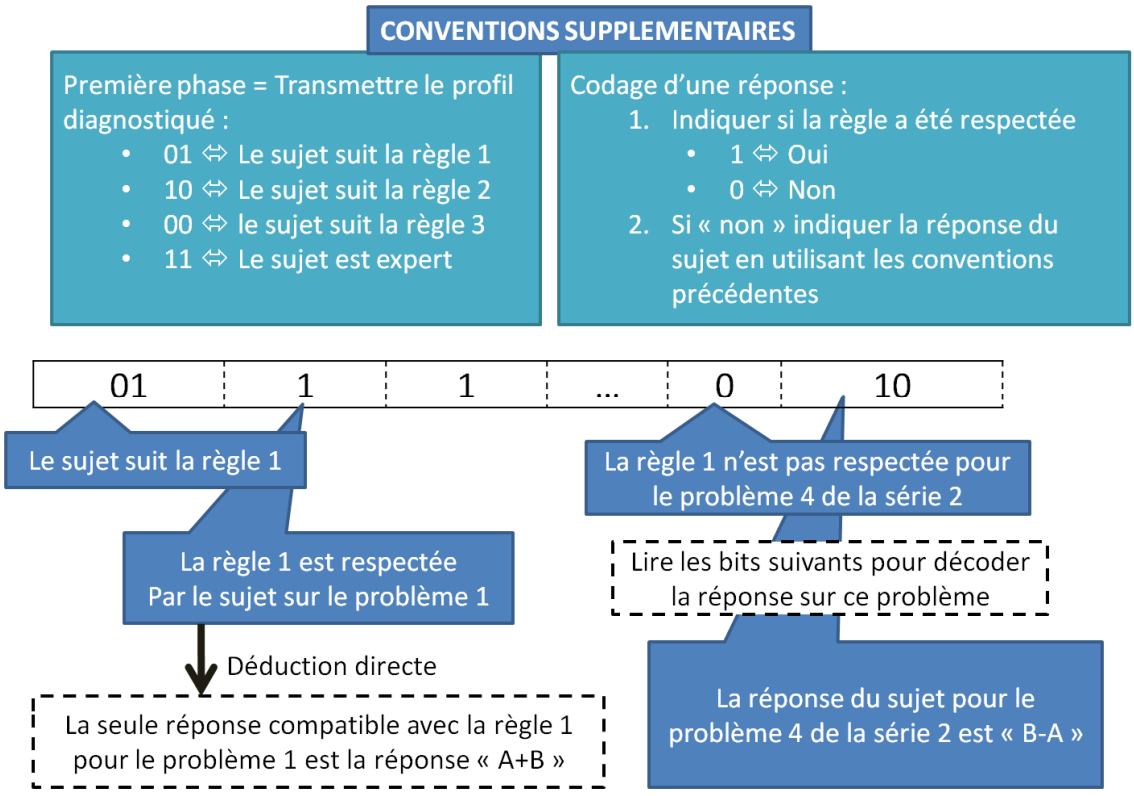



Figure 31. Transmission des réponses d'un sujet par un langage binaire.

Le codage peut être vu comme un message que l'émetteur cherche à transmettre au récepteur. Pour qu'il puisse être écrit et compris, un certain nombre de conventions sont « fixées à l'avance ». Cette analogie permet de bien souligner que toute l'information est contenue dans le message produit, sans perte ni ambigüité. Sa taille décrivant

l'ensemble des réponses de chaque sujet est donc de 16 problèmes x 4 séries x 2 bits = 128 bits.

9.1.2.3 Données compressées par le modèle

Maintenant, supposons que par le biais d'un module de diagnostic, le sujet 1 est diagnostiqué comme suivant la **règle 1**. Puisqu'elle réduit l'espace des possibles à une seule réponse, alors elle devrait logiquement permettre une stratégie de codage économique si elle est souvent respectée. La Figure 32 constitue une première proposition.

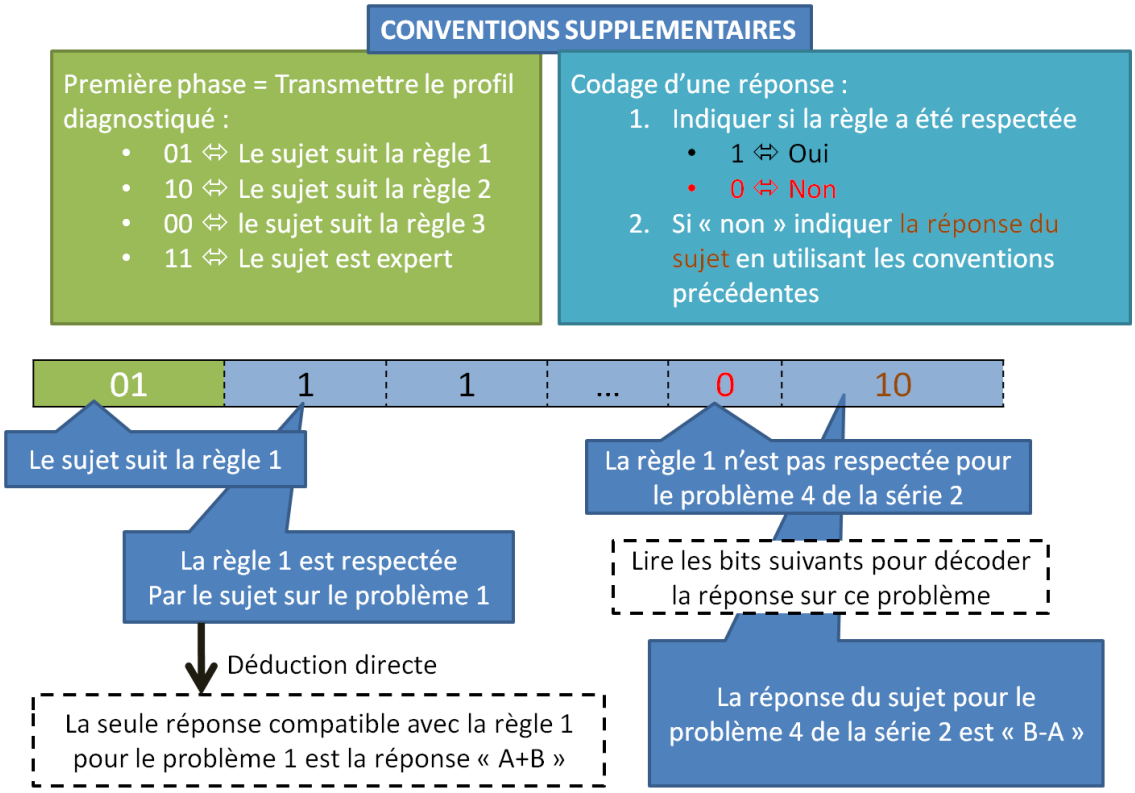



Figure 32. Transmission de la réponse d'un élève par le biais d'une convention de codage qui s'appuie sur le diagnostic de l'apprenant (en vert).

Deux nouvelles conventions sont nécessaires pour que l'émetteur et le récepteur puissent se comprendre. Il faut d'abord transmettre au récepteur quelle règle a été choisie pour représenter le sujet. Une convention a été construite à cet effet, le récepteur comprend donc que, par « 01 », le sujet est diagnostiqué par la règle 1. L'émetteur va ensuite déduire séquentiellement l'ensemble des réponses en regardant si pour chaque problème la règle a été respectée (0/1). S'il reste des informations nécessaires pour désambigüiser la réponse produite, alors le récepteur lit les binaires qui suivent pour savoir quelle est la réponse sélectionnée. Dans le cas présent, si elle suit la règle 1, il n'y

a pas d'information à rajouter, donc le récepteur connaît la réponse choisie et passe au problème suivant. Si la règle n'est pas respectée, il faut alors indiquer la réponse. La convention précédente est utilisée[42]. Il suffit alors de comparer la taille des deux messages pour savoir de combien de bits les données sont compressées.

Dans l'exemple de Tatsuoka, le sujet 1 respecte la règle 1 sur tous les problèmes à une exception près. La taille du message décrivant son protocole est donc de 2 bits (indiquant le diagnostic) + 63 x 1 bit + 1 x 3 bits = 68 bits ce qui revient à une compression de 128-68=60 bits.

### 9.1.3 Analyse et généralisation du critère

#### 9.1.3.1 Réalisation d'un compromis

Ce nouveau codage a un certain nombre de propriétés intéressantes qui semblent répondre à notre problématique. Si le nombre de règles possibles est grand, alors la première partie du code devient grande, ce qui revient à pénaliser le nombre de degrés de liberté du modèle. Problème par problème, si la règle diagnostiquée est respectée, alors la réponse de la taille du codage diminue (1 bit pour la règle 1), mais si ce n'est pas le cas, la taille augmente (3 bits). Par ailleurs, plus une règle est restrictive, plus elle permet de compresser. Par exemple, si le nombre de réponses possibles est de 8 et que la règle 1 ne prédit toujours qu'une réponse, alors le rapport grandit : la réponse est encodée par 1 seul bit si la règle est respectée, mais par 4 bits si elle ne l'est pas, contre 3 bits pour le codage normal.

Selon cet exemple introductif, la compression apparait être un cadre favorable pour exprimer des modèles non probabilistes, car elle établit automatiquement un compromis entre (1) le nombre de degrés de liberté du modèle, (2) la taille de l'étau (le nombre de

[42] Dans le cas où le profil n'est pas respecté, par déduction le nombre réel de réponses restantes n'est pas de 4 mais de « 3 ». Le codage peut donc être optimisé. Cependant dans le cadre de cet exemple, nous simplifions le cas et nous faisons appel à la convention utilisée pour décrire le message brut. Nous montrerons plus tard comment gérer ce cas.

réponses compatibles) et (3) le nombre de cas où le profil est erroné, c'est-à-dire lorsque la réponse sort de l'étau.

#### 9.1.3.2 Emploi du critère statistique

Le critère mis en place a plusieurs fonctions :

(1) comme valeur à maximiser : pour chaque élève, le profil cognitif qui offre la meilleure compression est sélectionné ;

(2) comme mesure de la fiabilité du profil cognitif associé à un élève ;

(3) pris en tant que moyenne sur plusieurs élèves, comme une quantité pouvant guider les choix de modélisation (ajout, suppression, réorganisation des règles, comparaison avec d'autres modèles) ;

(4) comme technique pour comparer des modèles qu'ils soient probabilistes ou non.

Ce dernier point est important, il touche au sens même des critères de sélection de modèles. Nous avons décrit une stratégie d'encodage d'une série de réponses d'élève qui ne dénature pas l'aspect non probabiliste du modèle. Comme nous l'avons souligné, des implémentations basées sur le principe du MDL existent aussi pour les modèles probabilistes, et une compression peut être calculée. Il est donc alors possible de comparer des modèles entre eux, quelle que soit leur nature. Ce n'est pas une problématique nouvelle, son intérêt est renouvelé par les récents travaux de Sébastien Lallé (Lallé, Luengo, & Guin, 2012; Lallé, Mostow, Luengo, & Guin, 2013), proches des nôtres. Une méthodologie consistant à projeter les modèles sur un même plan et à leur attribuer des prédictions probabilistes est proposée. Cette étape étant réalisée, ils peuvent être comparés en utilisant les critères les plus usuels en sélection de modèles (AIC, BIC).

Notre critère, outre la réalisation d'un compromis entre complexité et précision, fournit un outil pour remplir les objectifs que nous nous sommes fixés :

- Conservation de l'aspect non probabiliste du modèle.
- Évitement (partiel) des probabilités.
- Favorisant les modèles portant sur un degré de granularité fin.

Par ailleurs, il fait appel à un principe intuitif (minimum de complexité) et calculable facilement.

9.1.3.3 Possibilité de modéliser à un niveau plus fin de granularité

Une granularité plus fine n'implique pas par essence une perte de puissance statistique. Dans l'article « *Unified Cognitive/Psychometric Diagnostic Assessment Likelihood-Based Classification Techniques* » DiBello et collaborateurs (DiBello et al., 1995) ont ces propos (p. 365) :

> *« The key to the trade-off between adequate model verisimilitude and the statistical power needed to reliably gather useful cognitive information from simple tests lies in the control of "model granularity," that is of the level of attention paid by the model to small amounts of systematic variation in response behavior. In essence, the number of parameters increases greatly as the grain size becomes finer »*

Si cette idée est en effet vérifiée pour les Q-matrices, nous mettons en doute le caractère général de cette affirmation. Notre exemple précédent est en contradiction avec cette idée. Dans cette approche, le niveau de granularité fin n'est ni synonyme de perte de puissance statistique ni d'une augmentation du nombre de paramètres libres. Si le chercheur travaille avec des règles qui portent à un niveau plus fin que réussite/échec, alors la perte de puissance statistique va plutôt dans le sens inverse car elles offrent de meilleures possibilités de compression. Si nous pouvons concéder que l'ajout de règles dans un modèle augmente à la fois la granularité et le nombre de paramètres libres, la granularité d'un modèle peut être augmentée sans pour cela produire des nouveaux paramètres libres.

Le fait de considérer un ensemble de réponses possibles est alors ici une richesse. S'il est vrai qu'une approche psychométrique aurait tendance à multiplier ses degrés de liberté en augmentant le nombre de probabilités à estimer, ce n'est pas le cas pour une approche directement cognitive. La version que nous utilisons du MDL est très proche de l'analogie de compression de données, qui est un principe simple. Il pourrait correspondre à l'énoncé verbal suivant :

> *« Si nous voulons expliquer à un professeur ce qu'a fait l'élève par le biais d'un modèle, il ne faut pas qu'il y ait un ensemble trop grand de profils possibles, il est aussi désirable que lorsque l'élève suit un profil, son association rencontre peu d'exceptions et apporte un éclairage substantiel pour expliquer ses réponses. »*

9.1.3.4 Formulation plus générale

Jusqu'à présent nous avons présenté le critère uniquement sur un exemple introductif, nous en faisons maintenant une description plus générale. L'usage du modèle peut être double, il peut indiquer si un problème étant donné, la réponse de l'apprenant respecte le profil diagnostiqué. C'est ce que nous avons vu dans l'exemple précédent. Il peut aussi permettre d'affiner le codage de la réponse. En effet, dans l'exemple, la règle Numéro 1 ne prédisait qu'une seule réponse, nous pourrions imaginer un autre cas, dans lequel elle prédit deux réponses sur 10. Alors, si elle est respectée, il ne reste plus qu'à indiquer de quelle réponse il s'agit (1 bit), si elle n'est pas respectée, la réponse parmi les 8 réponses restantes doit être indiquée ce qui coûte 3 bits.

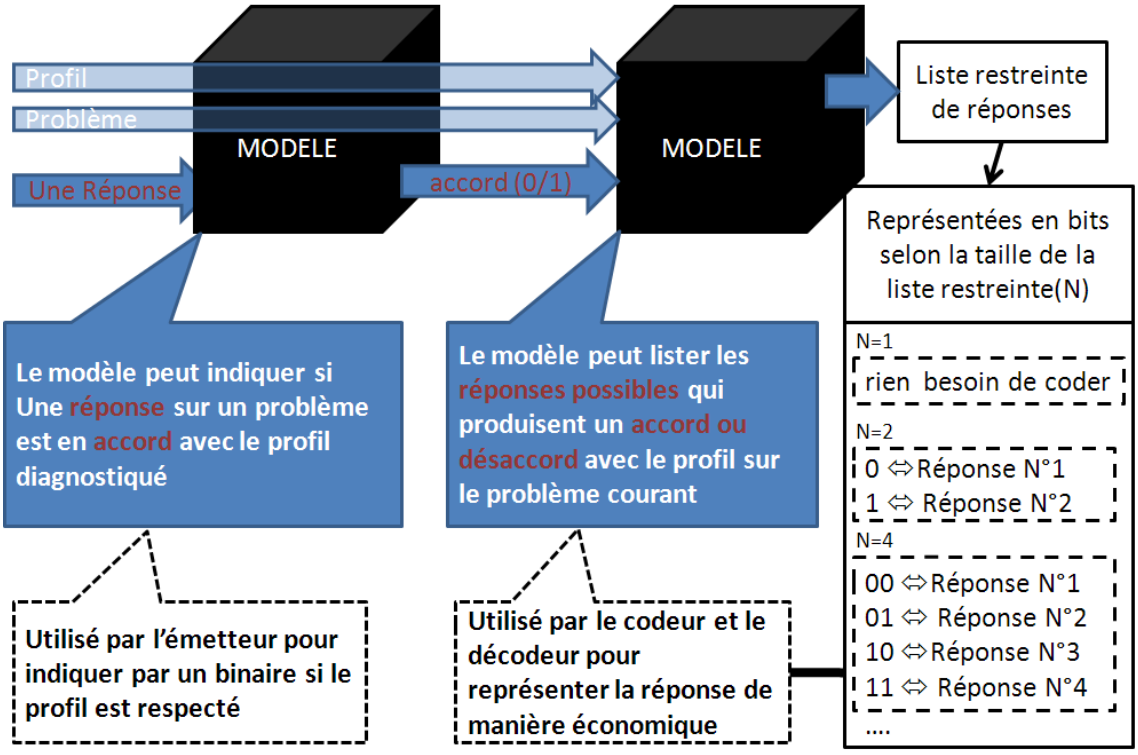



Figure 33. Illustration du double usage possible du modèle pour encoder ou décoder les réponses.

Il arrive, et c'est statistiquement le cas le plus fréquent, que la liste des réponses possibles ou que la liste des profils ne soit pas une puissance de 2. Ainsi, si 3 réponses sont possibles, la taille théorique du message est de $Log_2(3)=1.58$ bit. Cette longueur n'est pas purement théorique, car il est possible de mettre en place une stratégie de codage « globale » dont la taille du message est effectivement la somme de ses tailles « à virgules » (arrondie à l'unité supérieure cependant). Nous décrivons cette approche en annexe pour ne pas surcharger cette partie.

La Figure 34 synthétise le calcul de la taille « t » du message compressé dans le cas général.

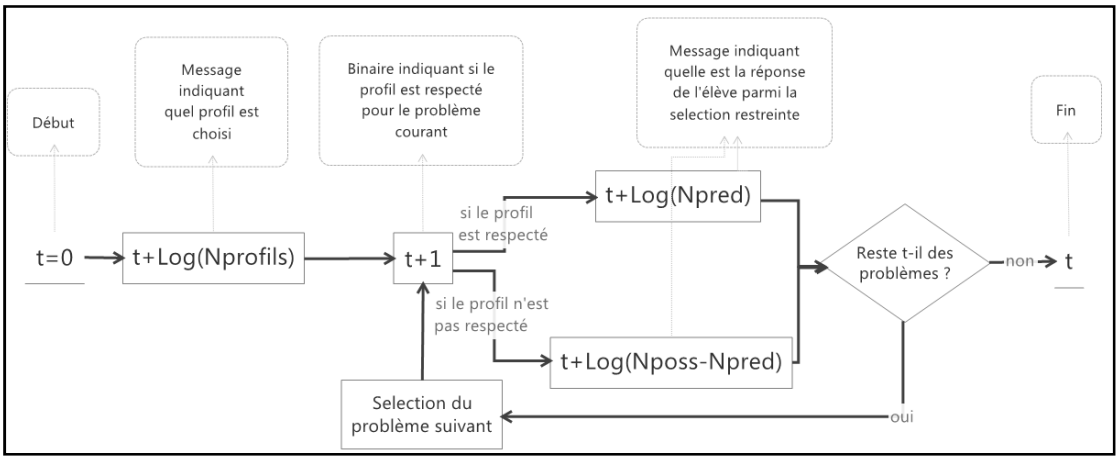



Figure 34. Algorithme de calcul de la taille du message compressé. Le log est en base 2.

La longueur totale peut alors être comparée à la taille du message décrivant sans modèle l'ensemble des réponses de l'élève. La taille est égale à la somme sur l'ensemble des problèmes des $Log_2$(Nposs) (voir Figure 35).

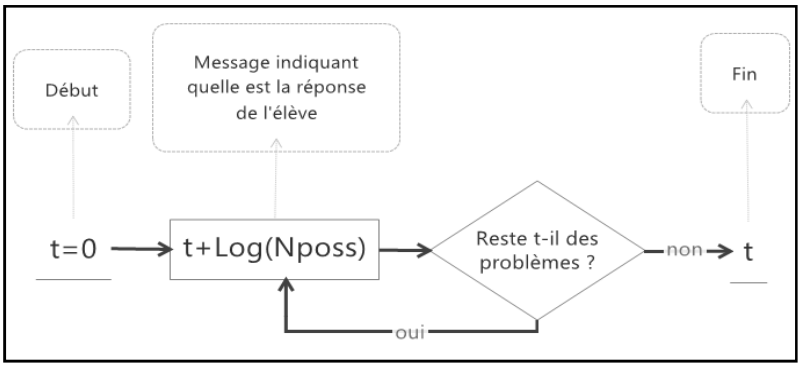



Figure 35. Calcul de la taille du message brut. Le log est en base 2.

## 9.1.4 Limite du critère si l'étau n'est pas assez resserré.

### 9.1.4.1 Description du problème et des premières solutions

Le critère développé est basé sur l'idée qu'un modèle non probabiliste qui ne se trompe pas apporte une réduction de complexité. Pour que ce choix soit pertinent, **il faut que le nombre de réponses prédites soit petit devant le nombre de réponses possibles**. Cela définit un cadre d'utilisation du critère qu'il convient de souligner. En reprenant les termes précédents si le rapport $r = \frac{Npred}{Nposs}$ se rapproche de $\frac{1}{2}$, alors le critère est à éviter dans sa forme brute. Il est possible de montrer mathématiquement que lorsque

$r = \frac{1}{2}$, le critère n'apporte aucune réduction de complexité quel que soit les réponses observées. En effet, si le profil est respecté, la réponse de l'élève se situe parmi les réponses non prédites, c'est-à-dire dans la deuxième moitié des réponses possibles. Le profil n'a donc pas plus d'intérêt à être respecté qu'à être non respecté. Face à ce problème, nous avons plusieurs solutions possibles :

La première est de ne coder avec le modèle que les réponses pour lesquelles $r < \frac{1}{2}$. Il suffit d'adapter la décision de ne pas utiliser le modèle pour coder les réponses lorsque les profils ne restreignent pas assez l'ensemble des réponses[43]. Cette méthode a l'avantage d'être simple et de permettre de conserver l'aspect déterministe du modèle. En effet, si nous devions verbalement décrire les réponses d'un sujet, il peut être compréhensible de n'utiliser le profil de diagnostic que dans les cas où il réduit au moins de moitié l'ensemble des réponses possibles. Cependant cette méthode constitue un appauvrissement statistique du modèle. En effet, elle évite les données qui peuvent pourtant avoir un poids dans la validation du modèle.

Les autres alternatives consistent à coder sur moins de bits les réponses prédites des réponses non prédites, ainsi le modèle est favorisé s'il émet souvent des prédictions justes. Cela revient à faire intervenir des probabilités pour optimiser l'encodage de l'information.

#### 9.1.4.2 Solution choisie : Emploi modéré des probabilités

La deuxième solution est de remplacer le binaire indiquant si le profil est respecté ou non par un code plus pertinent. Par exemple, supposons que tout profil a 3 chances sur 4 d'être respecté, alors par un calcul d'entropie il est possible d'utiliser un code différent si le profil est juste (de longueur $Log_2(1/(¾))$ ) ou faux (de longueur $Log_2(1/(¼))$ ). Dans le cas présenté, si un profil a effectivement 3 chances sur 4 d'être respecté, la longueur théorique espérée est de $1/4*( Log_2(1/ (¼) ) + 3/4*( Log_2(1/ (¾))) = 0.81$ bit, ce qui représente un léger gain par rapport au coût initial (d'un bit). Des algorithmes,

[43] La réponse est alors codée de la même manière que dans la version sans modèle, donc avec un coût de $Log_2(Nposs)$ bit.

comme celui de Huffman, nous garantissent de trouver une stratégie de codage qui tire parti de ces probabilités sans trop s'éloigner de la taille idéale espérée. Par contre, plus la probabilité réelle s'éloigne de celle espérée, moins le gain sera grand. Par exemple, si « dans la réalité » tout profil a 2 chances sur 3 d'être respecté (au lieu de la prévision à ¼), alors le calcul donne une longueur espérée de 0.94 bit. Cette technique a aussi ces limites, notamment lorsque le rapport dépasse la valeur seuil de 1/2. En effet si le rapport $r = \frac{Npred}{Nposs}$ vaut 1-p avec p, la probabilité que le profil soit respecté alors il est possible de montrer dans le cas général qu'aucune compression n'est réalisée. Intuitivement, si $r = \frac{Npred}{Nposs} = 3/4$, alors avoir un modèle qui fait 3 prédictions justes sur 4 n'est pas meilleur que le hasard. La troisième solution consiste à combiner les deux approches précédentes. La réponse n'est codée avec le modèle que si r est inférieur à 3/4. Elle permet donc de limiter la perte de données causée par le problème 1. Cette approche supprime néanmoins des données qui peuvent malgré tout avoir de l'intérêt.

Une dernière possibilité, celle que nous retenons, est de coder cette information de manière **dynamique** en considérant que le modèle est k-fois meilleur que le hasard. Nous pouvons par exemple considérer que le modèle est deux fois meilleur que le hasard. Supposons que $\frac{Npred}{Nposs} = 2/3$. Si nous prenons k=2, alors le modèle a une chance sur 6 seulement de se tromper (contre une chance sur 3 ce qui correspondrait au hasard). Ce rapport évite les risques d'incohérence du critère, ce qui est un point essentiel. C'est donc l'option que nous choisissons.

Cette technique peut être critiquée puisqu'elle emploie une approche probabiliste que nous souhaitions éviter. Cependant elle est absolument minimale, car elle porte sur la résultante d'un ensemble de règles (et non pas sur les règles elles-mêmes). Par ailleurs les probabilités utilisées ne consomment aucun degré de liberté, car elles sont calculées dynamiquement. Elles forment donc la bordure du critère et non pas son cœur, ce qui permet au chercheur de ne pas se préoccuper de cet aspect dans son activité de modélisation. Par ailleurs, il est intéressant de noter qu'au plus l'étau est fort, moins ce coefficient a de l'importance dans le calcul de compression.

## 9.1.5 Meilleure mise en évidence des différences interindividuelles

### 9.1.5.1 Définition du problème

Nous demandons au lecteur de considérer ces deux problèmes proches :

- La règle R, qui élimine une partie des réponses, est pratiquement toujours diagnostiquée chez les sujets. Cela signifie que le processus de solution est mieux compris. Cependant la règle R ne contribue pas à une meilleure connaissance des différences interindividuelles et pourtant peut améliorer nettement le taux de compression du modèle.
- La liste de réponses attendues contient des réponses très peu fréquentes. Deux cas peuvent alors se produire :
    - Les règles généralement diagnostiquées les excluent et le modèle est artificiellement favorisé.
    - Les règles ne les excluent pas et le modèle est artificiellement défavorisé.

L'organisation des règles est aussi un aspect critique de notre évaluation. Supposons qu'une règle puisse être exprimée comme proche de l'union de deux autres règles du modèle. Si ce cas se produit, alors la règle générale a très peu de chance d'être sélectionnée. Or, rajouter des règles qui ont peu de chance d'apparaitre dans les profils diagnostiqués ou avoir des règles mal organisées augmente artificiellement la taille du code préfix indiquant quel profil a été sélectionné. Pourtant, il est tout à fait possible d'imaginer que les règles sont pertinentes et décrivent bien les différences interindividuelles.

En résumé, selon la manière dont le modèle ou les données sont construits, il est possible d'observer des compressions alors que les différences interindividuelles ne sont pas bien cernées et réciproquement. L'efficacité globale du modèle doit donc être distinguée de sa capacité à mettre en évidence de manière fiable des différences interindividuelles.

### 9.1.5.2 Test de permutation pour déterminer si les différences interindividuelles sont capturées par le modèle

Nous définissons un élève « de Frankenstein » comme un élève virtuel, construit par tirage aléatoire de ses réponses sur les différents problèmes. Ce tirage n'est pas uniforme, il se base sur les réponses que nous avons des autres élèves. Ainsi, si Jean et Fatma font partie de notre échantillon, un élève « de Frankenstein » pourrait avoir pour

réponse à la question 1 la réponse de Jean, et la réponse de Fatma pour la question 2. Pour chaque expérimentation, il existe $n^p$ possibilités d'élèves « de Frankenstein » avec p valant le nombre de problèmes plus 1 et n représentant le nombre d'élèves de l'échantillon considéré.

Chacun de ces élèves est diagnostiqué et le score basé sur le MDL est calculé. Comme l'ensemble des problèmes listés précédemment affecte également les élèves, c'est à dire complètement indépendamment de leurs différences interindividuelles, alors il affecte de la même manière les élèves « de Frankenstein » et réels. Une autre propriété de ces élèves virtuels est que, **sous l'hypothèse nulle selon laquelle le modèle ne capture pas les différences interindividuelles, alors l'évaluation de leur diagnostic ne devrait pas être significativement moins bonne que celle des élèves réels**.

Réciproquement, si la différence entre le score d'un élève réel et la moyenne des scores des élèves « de Frankenstein » est nette, alors nous avons un bon argument pour supporter l'idée que notre modèle capture des différences interindividuelles. Cela ne signifie pas que toutes les dimensions du profil sont pertinentes, mais que le profil dans son ensemble, capture ce qui distingue l'individu du groupe.

De manière plus quantitative, à un élève diagnostiqué, il est possible de compter la proportion des cas pour lesquels un élève « de Frankenstein » a un meilleur diagnostic par de multiples tirages. La Figure 36 illustre le fonctionnement de ce test. Le principe du test de permutation est le même que celui utilisé dans la partie sur la génération d'erreur. Il consiste à établir une distribution des permutations sur notre statistique d'intérêt. Le ratio calculé est équivalent à un calcul de p-value. Dans cette partie, les prédictions des modèles étaient permutées et la statistique d'intérêt était le nombre d'occurrences captées par le modèle. Ici, ce sont les données qui sont permutées et notre statistique d'intérêt devient le calcul du taux de compression des modèles produits.

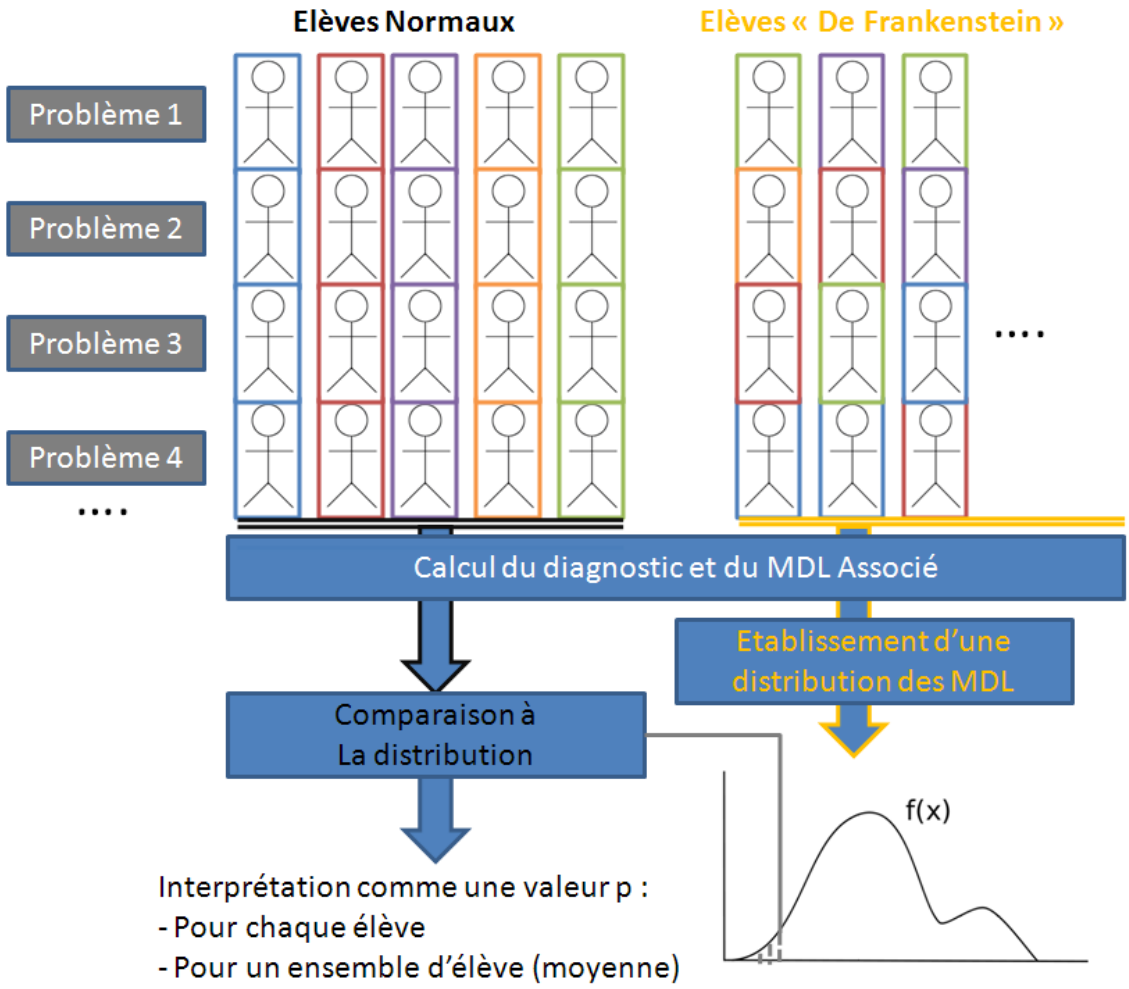



Figure 36. Schéma représentant l'analyse portée par un test de permutation

Il est facile de voir, en reprenant le calcul du critère, que l'ajout d'une règle ou d'une réponse inutile affecte autant les élèves normaux que les élèves artificiels.

Du fait de ces propriétés intéressantes, l'usage de ce test met-il l'usage du MDL en valeur absolue au second plan ? Les deux indicateurs sont porteurs d'informations complémentaires, comme le présente le Tableau 18.

Tableau 18. Évaluation d'un modèle par le biais d'une mesure basée sur le MDL et d'un test de permutation.

| | | Compression | |
|---|---|---|---|
| | | non | oui |
| Différence avec les élèves "De Frankenstein" | faible ou nulle | BILAN : Negatif<br>RECOMMENDATION : Revoir le modèle dans son ensemble. | BILAN : Mixte. Le modèle représente bien les sujets d'un point de vue général, mais pas d'un point de vue interindividuel.<br>RECOMMENDATION : Ajouter des règles cherchant à établir des traits individuels (ex : compétences) |
| | forte | BILAN : Mixte. Le modèle semble capturer les différences interindividuelles, mais peine à les représenter de manière synthétique.<br>RECOMMENDATION : Tester la suppression ou la réorganisation des règles pour améliorer la compression. Augmenter la taille des données. | BILAN : Positif<br>RECOMMENDATION : Aucune recommendation spécifique. |

Cette technique utilisée pour capturer les différences interindividuelles peut être mise en regard avec une proposition similaire donnée lors de la présentation d'ASPM (Simon et al., 1995). Les auteurs suggèrent de tester le modèle sur la réponse modale des sujets (la plus fréquente) pour comparer ses performances avec celles obtenus sur des sujets normaux et ainsi tester si des différences interindividuelles sont mises en évidence. Nous argumentons que le test de permutation est une technique plus puissante pour cet objectif du fait de l'extraction d'une distribution plutôt qu'une unique valeur. En effet, supposons que le modèle fait un excellent diagnostic sur le sujet reconstitué par les réponses modales, qu'en déduire, sinon que « *nous n'avons pas eu de chance* » ? De plus, ce cas a de forts risques de se produire si la réponse la plus populaire est la réponse juste, le modèle expert serait alors parfaitement prédictif.

## 9.2 STAR : Outil auteur pour la conception et l'évaluation de modèle cognitif

Notre objectif est d'offrir un espace d'écriture et d'évaluation aux modèles symboliques simples. Mettre en place une mesure quantifiant la pertinence des diagnostics réalisés a été un premier pas dans cette direction. Le but est de permettre au modélisateur de construire son modèle de manière exploratoire. En effet, dû à leur complexité et à la difficulté de la tâche, les modèles symboliques ne se construisent généralement pas en « un coup ». Dans les travaux portant sur la mise en place du modèle des contraintes, par exemple, la modélisation est une activité itérative supervisée par le chercheur (Richard et al., 2009, 1993). Ainsi, modéliser par les contraintes est plus complexe qu'une estimation de paramètres dans un modèle probabiliste. La complexité de la tâche

de modélisation entraîne des coûts en temps important, en témoigne les remarques conclusives dans (Richard et al., 2009, p. 21):

> *La dernière remarque concerne l'intérêt de la modélisation des processus individuels de résolution. [...] Même si cette approche peut apporter des informations complémentaires, on peut se demander si elle justifie l'investissement nécessaire.*

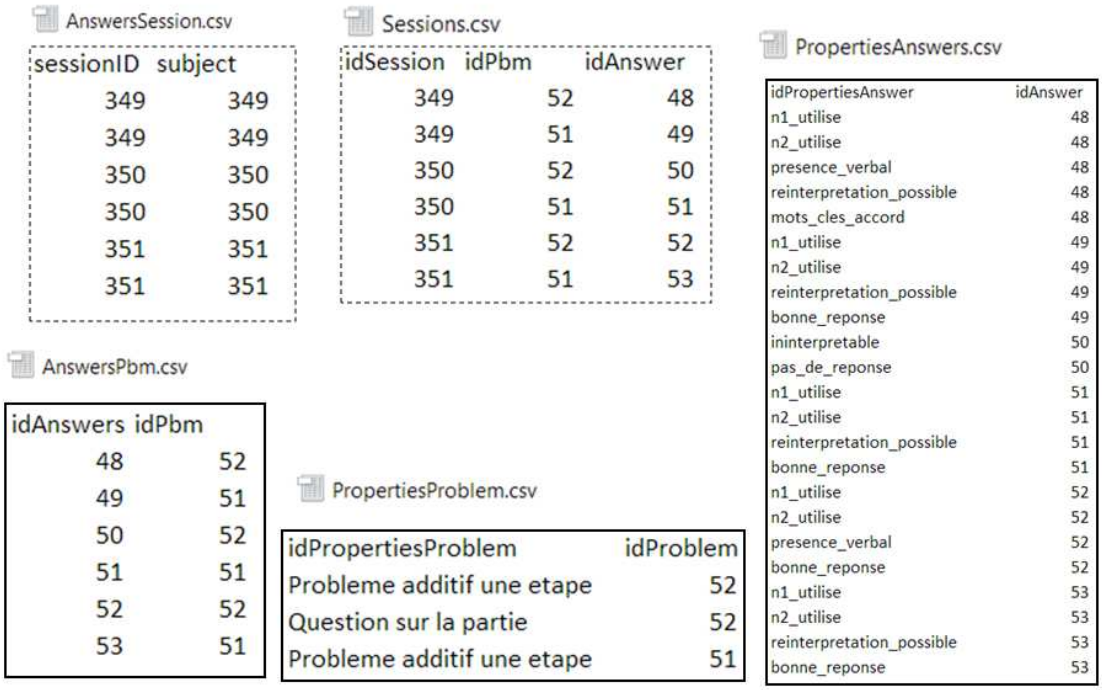
Permettre la modélisation rapide avec un outil adapté permettrait alors de gagner du temps lors de la conception d'un modèle cognitif, ce qui pourrait encourager de futures modélisations. Nous présentons dans cette partie un outil minimal qui répond à cette problématique dans les environnements d'apprentissage. Il peut être directement testé sur les données qui lui sont fournies et calculer l'indicateur que nous avons mis en place.

## 9.2.1 Le fonctionnement du programme

L'interface de STAR est constituée d'onglets faisant chacun appel à une bibliothèque de fonctions. Si l'utilisation de cette bibliothèque est réservée au développeur, l'interface graphique permet d'élargir STAR à un public non formé à la programmation. Les trois premiers onglets ouvrent, manipulent et enregistrent des fichiers qui leur sont propres. Le dernier est le plus important, car il calcul le critère précédemment décrit. Dans la construction du modèle général, nous proposons une structure en trois niveaux que nous décrivons par le fil conducteur d'une analogie basée sur la matière.

### 9.2.1.1 Données : Niveau sous-atomique

La première couche est celle du matériel élémentaire, constituée des données exportées de l'EIAH. Par convention, l'extension de l'archive porte le nom .dataP pour « data problem ». Ces dernières décrivent, par le biais de simples tables (cf. Figure 37), les informations essentielles sur les sessions d'exercices réalisées par les enfants.



AnswersSession.csv

| sessionID | subject |
|---|---|
| 349 | 349 |
| 349 | 349 |
| 350 | 350 |
| 350 | 350 |
| 351 | 351 |
| 351 | 351 |

Sessions.csv

| idSession | idPbm | idAnswer |
|---|---|---|
| 349 | 52 | 48 |
| 349 | 51 | 49 |
| 350 | 52 | 50 |
| 350 | 51 | 51 |
| 351 | 52 | 52 |
| 351 | 51 | 53 |

PropertiesAnswers.csv

| idPropertiesAnswer | idAnswer |
|---|---|
| n1_utilise | 48 |
| n2_utilise | 48 |
| presence_verbal | 48 |
| reinterpretation_possible | 48 |
| mots_cles_accord | 48 |
| n1_utilise | 49 |
| n2_utilise | 49 |
| reinterpretation_possible | 49 |
| bonne_reponse | 49 |
| ininterpretable | 50 |
| pas_de_reponse | 50 |
| n1_utilise | 51 |
| n2_utilise | 51 |
| reinterpretation_possible | 51 |
| bonne_reponse | 51 |
| n1_utilise | 52 |
| n2_utilise | 52 |
| presence_verbal | 52 |
| bonne_reponse | 52 |
| n1_utilise | 53 |
| n2_utilise | 53 |
| reinterpretation_possible | 53 |
| bonne_reponse | 53 |

AnswersPbm.csv

| idAnswers | idPbm |
|---|---|
| 48 | 52 |
| 49 | 51 |
| 50 | 52 |
| 51 | 51 |
| 52 | 52 |
| 53 | 51 |

PropertiesProblem.csv

| idPropertiesProblem | idProblem |
|---|---|
| Probleme additif une etape | 52 |
| Question sur la partie | 52 |
| Probleme additif une etape | 51 |

Figure 37. Exemple de cinq tables contenues dans l'archive d'extension .dataP. Les tables en pointillés donnent des informations sur les réponses des élèves et celles aux traits pleins renseignent les propriétés des problèmes et des réponses.

Chaque problème est étiqueté d'un certain nombre de propriétés « problème » et chaque réponse d'apprenant est étiquetée de propriétés « réponses ». L'exemple le plus simple de propriété « réponse » est « bonne_reponse ». Si le niveau de la modélisation est plus fin, il est possible de décrire des types d'erreurs par l'utilisation de ces propriétés. Dans STAR, le premier onglet permet de charger une archive contenant les fichiers présentés. Lors de l'ouverture, l'utilisateur se voit demander s'il souhaite que les données soient enrichies par les propriétés « opposées ». Cette étape permet de flexibiliser l'écriture des règles. Ainsi, si un problème n'a pas la propriété « bonne_réponse », alors il détient la propriété « ABS_OF_bonne_reponse » (cf. Figure 38).

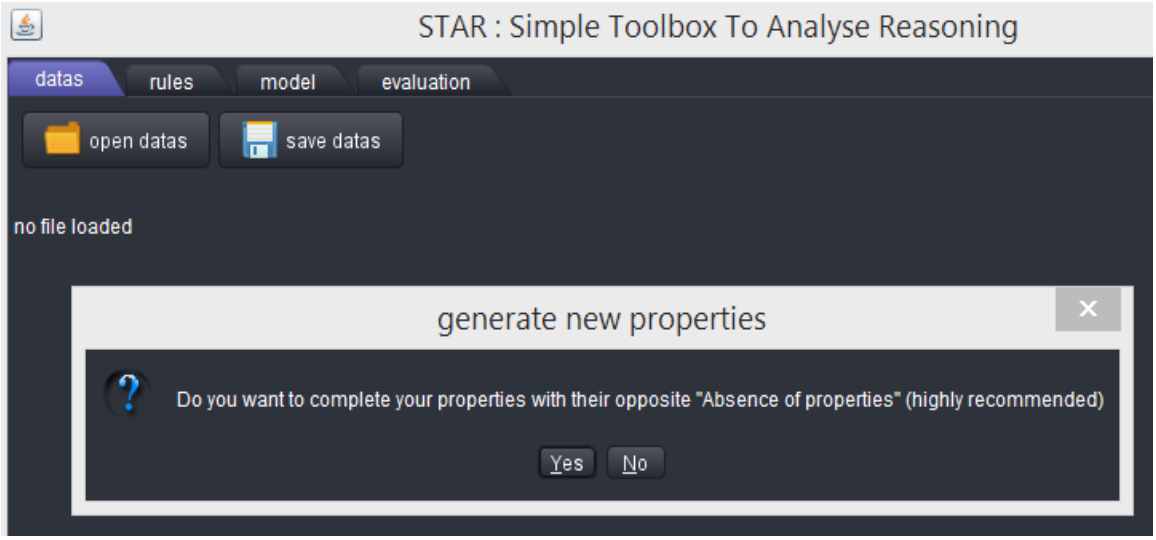



Figure 38. Message à l'ouverture d'une archive dataP contenant les données à analyser dans STAR.

9.2.1.2 Rules : Le Niveau atomique

Le niveau atomique porte sur la construction de règles. Le terme « règle » est à prendre au sens large, car elles peuvent représenter des contraintes ou des compétences. Leur écriture est relativement simple, elles sont formalisées comme une matrice de binaires. Ces dernières ont pour colonnes les propriétés des problèmes et pour lignes les propriétés des réponses. Elles décrivent donc une formule logique entre les propriétés réponses et problèmes.

Une règle peut être présente ou non chez l'apprenant. Les détecter est l'objet même du diagnostic. Elles se lisent comme *« si l'élève a la règle R ci-décrite, alors il ne donne pas de réponses qui ont la propriété P_r à des problèmes qui ont la propriété P_p »*. Les règles dans STAR se rapprochent préférentiellement de la notion de contrainte.

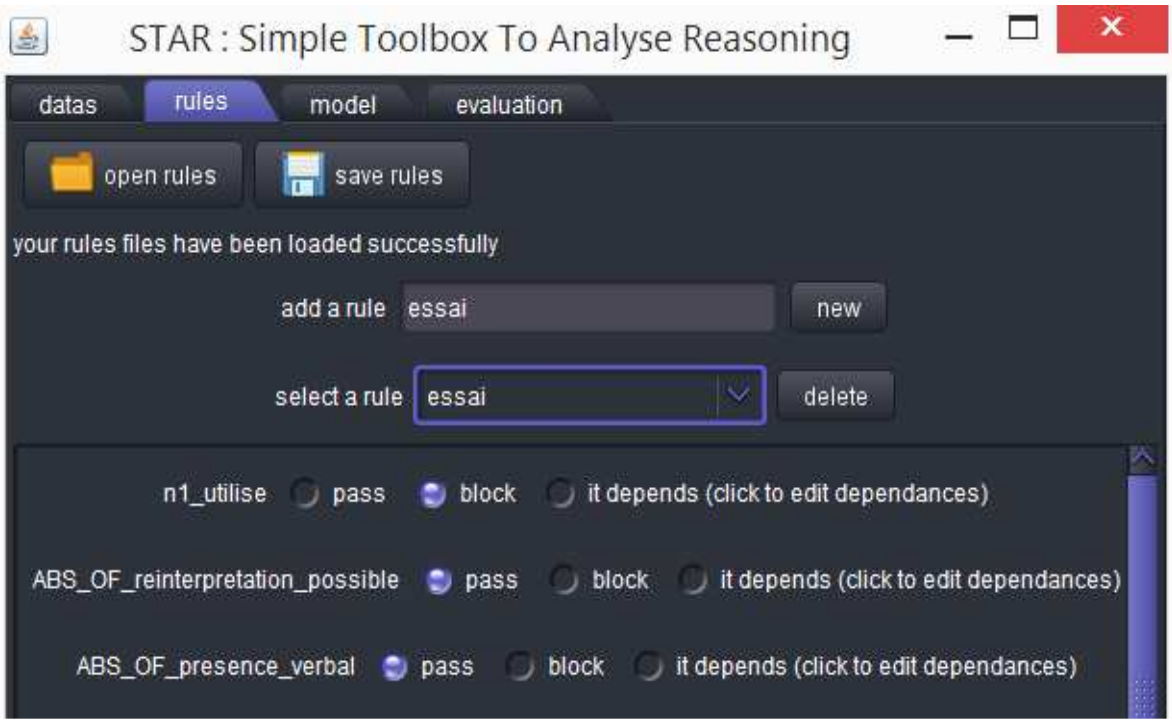



Figure 39. Onglet de définition de règles dans STAR. Dans cet exemple, la règle « essai » interdit la propriété réponse « n1_utilise ».

Par défaut, les règles ne portent que sur les propriétés « réponse ». Elles sont donc de la forme simplifiée *« si l'élève a la règle R ci-décrite, alors il donne des réponses qui ont la propriété R »*. Si l'utilisateur le souhaite, il peut faire des distinctions en utilisant les propriétés problèmes, c'est ce que montre la Figure 40. Dans ce cas, l'utilisateur peut indiquer quelles cellules de la matrice [propriété réponse] x [propriétés problème] sont autorisées. Dans cet onglet, il est aussi possible de sauvegarder et de charger ces règles en fichier zip, d'extension .dataR pour « data Rules ». Cette archive contient l'ensemble des matrices représentant ces règles.

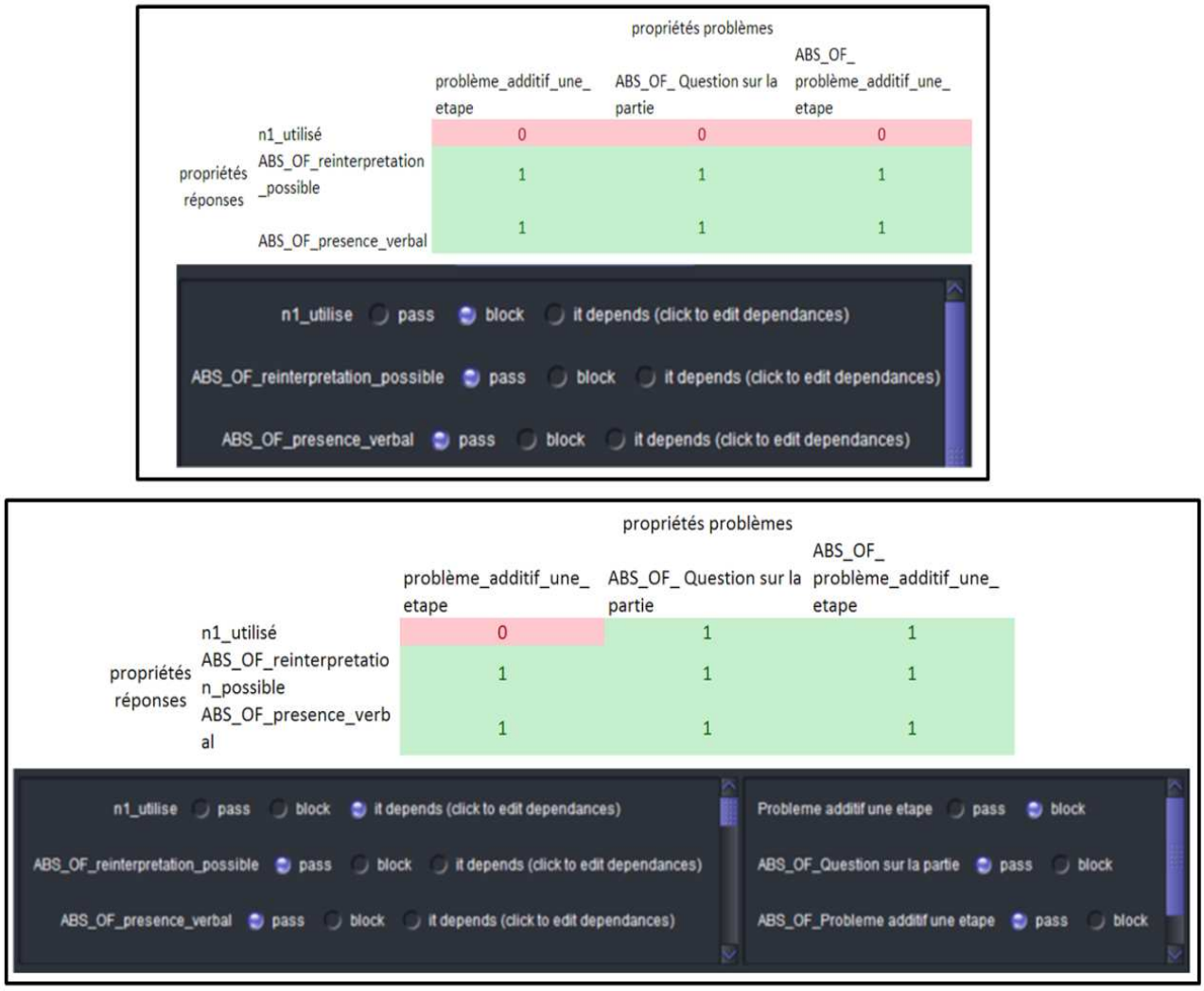



Figure 40. Définition des règles dans STAR. La figure est constituée de deux vues tronquées de l'interface de création de règles dans STAR et associée à l'équivalent en terme de matrice des options sélectionnées. Par défaut une règle ne porte que sur les propriétés-réponses (premier encadré), il est possible de préciser les propriétés problèmes en cliquant sur « it depends » (deuxième encadré).

Lorsque des règles sont contradictoires, nous considérons que le « non l'emporte ». La loi de composition des règles dans STAR est par défaut l'accumulation des exclusions. Par exemple si nous avons :

- La règle1 déclare que les réponses sur l'ensemble des problèmes ont la propriété 1 et 2
- La règle2 déclare que l'ensemble des réponses a la propriété 2 et 3.

**Alors,** si un élève est supposé suivre les règles 1 **et** 2 il est considéré que ses réponses auront seulement la propriété 2.

9.2.1.3 Modèle : Niveau moléculaire

Une molécule peut être composée d'un seul ou de plusieurs atomes à condition que sa structure soit stable. Dans cet esprit, le troisième onglet permet de définir les différents assemblages possibles, c'est-à-dire l'organisation des règles entre-elles.

Il est difficilement imaginable de concevoir un système absolument général pour l'organisation des règles. Dans certains cas, le diagnostic est composé d'une unique règle, c'est par exemple le cas dans le Rule Space Model de Tatsuoka (1983). Dans d'autres cas, comme pour le modèle des contraintes, elles sont organisées en familles (Richard, Pastré, & Parage, 2009). Un diagnostic est construit en choisissant une unique contrainte par famille. Cette technique permet de forcer une sélection de contraintes lorsqu'elles portent sur un même aspect du modèle cognitif.

Dans STAR, l'organisation des règles entre elles est traduite par une matrice d'exclusion. Cette dernière est carrée et symétrique, constituée de binaires indiquant si la règle i est compatible avec la j. Elle permet d'exprimer des modèles divers, comme le présente la Figure 41. Sa diagonale indique les règles testées dans le modèle général. Lorsqu'elle contient un zéro, alors la règle s'auto-exclut et est écartée du modèle.

| | a | b | c | d |
|---|---|---|---|---|
| a | 1 | 0 | 0 | 0 |
| b | 0 | 1 | 0 | 0 |
| c | 0 | 0 | 1 | 0 |
| d | 0 | 0 | 0 | 1 |

| | a | b | c | d |
|---|---|---|---|---|
| a | 1 | 1 | 1 | 1 |
| b | 1 | 1 | 1 | 1 |
| c | 1 | 1 | 1 | 1 |
| d | 1 | 1 | 1 | 1 |

| | a | b | c | d |
|---|---|---|---|---|
| a | 1 | 0 | 1 | 1 |
| b | 0 | 1 | 1 | 1 |
| c | 1 | 1 | 1 | 0 |
| d | 1 | 1 | 0 | 1 |

| | a | b | c | d |
|---|---|---|---|---|
| a | 1 | 1 | 1 | 1 |
| b | 1 | 1 | 1 | 1 |
| c | 1 | 1 | 0 | 1 |
| d | 1 | 1 | 1 | 0 |

Figure 41. Différentes organisations d'un ensemble de quatre règles. De droite à gauche : (1) Modèle dans lequel seule une règle est sélectionnée, car chacune est en exclusion avec toutes les autres. (2) Modèle sans exclusion inter-règle (toutes les combinaisons de règles sont possibles pour former un diagnostic). (3) Modèle dans lequel {a,b} et {c,d} sont réparties en familles. (4) Modèle complet dans lequel les règles {c} et {d} sont exclues de l'analyse.

L'interface pour produire cette matrice est plutôt minimale, les deux premiers modèles présentés dans la Figure 41 sont accessibles en un clic par des boutons dans l'interface présentée en Figure 42.

Dans le cas où l'utilisateur veut produire un modèle plus complexe il doit écrire la matrice lui-même et la charger dans STAR. Le programme essaye alors d'ouvrir un fichier csv (un tableau représentant la matrice) avec le programme qui y est associé (comme Excel, Open Office, etc.). Si aucun logiciel n'est associé, alors l'utilisateur se

voit demander d'ouvrir manuellement ce fichier, en sélectionnant lui même son tableur. Lorsque l'utilisateur a complété sa matrice, il la charge dans le programme. Étant donné que cette étape peut être fastidieuse, nous avons offert la possibilité d'une construction automatique de matrices.

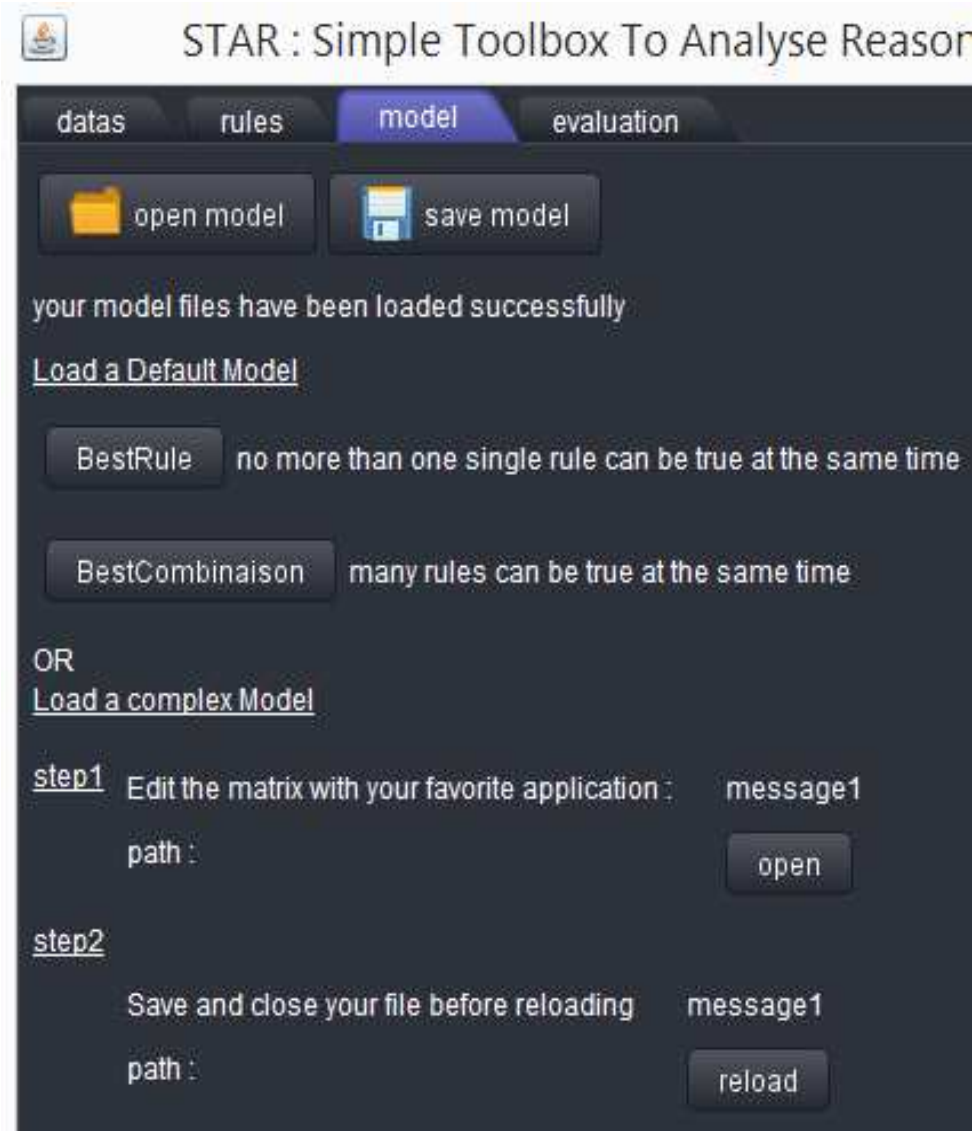



Figure 42. Construction d'un modèle dans STAR.

Un mode automatique est possible par la construction de modèles simplifiés (Best Rule choisit la meilleure règle, et BestCombinaison choisit le meilleur assemblage de règle). Si l'utilisateur veut concevoir des modèles plus complexes, il peut éditer et charger la matrice avec un logiciel de type tableur avec les boutons « open » et « reload ».

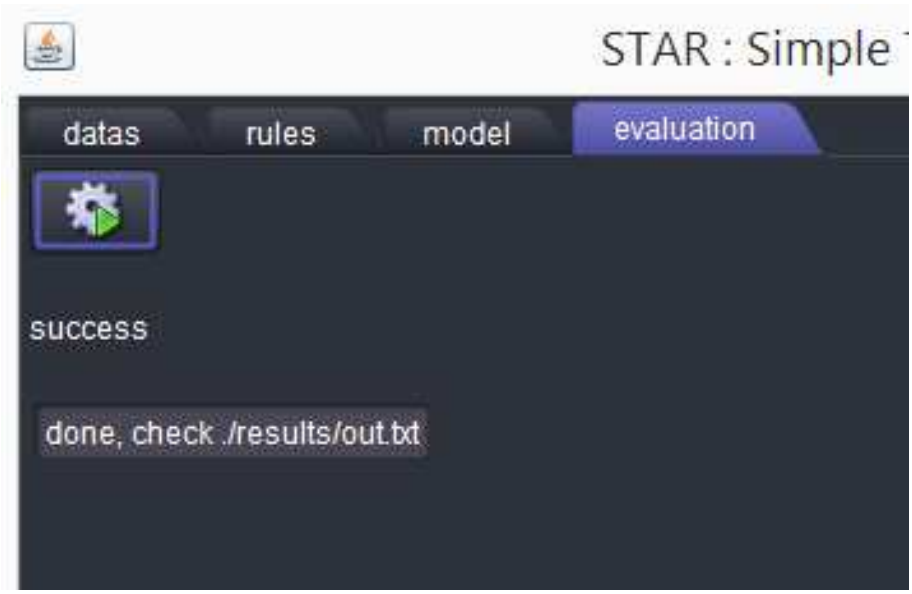
Une molécule peut être composée d'un seul atome, il est donc possible de tester la pertinence d'une seule règle.

La matrice correspondante, donc, ne contient qu'un seul 1 dans sa diagonale, à l'endroit correspondant à la règle testée.

La représentation matricielle dans l'onglet règles et l'onglet modèle n'est pas seulement une vue de l'esprit. Lorsque l'utilisateur sauvegarde son travail, des archives sont construites contenant des fichiers textes décrivant les matrices qui ont été réalisées par le biais de l'interface.

#### 9.2.1.4 Évaluation du modèle général

Le dernier onglet permet de quantifier la pertinence du modèle construit. Il permet de calculer la valeur pour quantifier la pertinence du diagnostic de chaque élève. En plus des mesures de complexité, un test de permutation est calculé pour compléter l'analyse. L'ensemble de ces valeurs et du résultat des diagnostics sont stockés dans un fichier externe à STAR (cf. Figure 43).




| Subject | modelSize | noModelSize | pval | diagnostic | problems |
|---|---|---|---|---|---|
| Subject1 | 28,68 | 25,25 | 0,8758 | noReuseNumbers,conceptualAbility_Level2, | 1_2_3_4_5_6_7_8_10_ |
| Subject2 | 27,97 | 25,25 | 0,7241 | noReuseNumbers,conceptualAbility_Level1, | 1_2_3_4_5_6_7_8_10_ |
| Subject3 | 27,97 | 25,25 | 0,7241 | noReuseNumbers,conceptualAbility_Level1, | 1_2_3_4_5_6_7_8_10_ |
| Subject4 | 23,46 | 25,25 | 0,0639 | conceptualAbility_Level3, | 1_2_3_4_5_6_7_8_10_ |
| Subject5 | 26,76 | 28,25 | 0,0829 | noReuseNumbers,conceptualAbility_Level3, | 1_2_3_4_5_6_7_8_9_10_ |

Figure 43. Dernier onglet de STAR, permettant de lancer les analyses. Au-dessous : forme du fichier de sortie (out.txt), indiquant la taille des données avec ou sans modèle, la p-value du test de permutation portant sur les différences interindividuelles, le diagnostic et la liste des problèmes sur lequel chaque sujet est passé.

## 9.2.2 Portée et limite du programme

### 9.2.2.1 Interopérabilité dans les EIAH

Précédemment, nous avons choisi de présenter STAR par son interface utilisateur plutôt que par les bibliothèques de fonction constituées. Elles sont au nombre de trois :

- STAR_Description : Lit et organise les informations contenues dans l'archive dataP. Cette bibliothèque est donc liée au premier onglet de l'interface graphique présentée.
- STAR_Modelisation : Charge, sauvegarde et modifie l'ensemble des informations liées à la construction du modèle. Cette bibliothèque est donc liée aux deux onglets suivants dans STAR.
- STAR_Evaluation : Biblothèque liée au dernier onglet, chargée d'établir le diagnostic de chaque élève sur la base du modèle et des données. C'est la plus lourde, car un certain nombre d'étapes sont nécessaires à l'établissement d'un diagnostic :
    - Lister l'ensemble des assemblages de règles possibles.
    - Construire les prédictions de cet assemblage.
    - Évaluer la pertinence de ces prédictions.

Du fait de la barrière des différents langages de programmation, il est difficile de développer des modules qui puissent être intégrés dans d'autres EIAH. Plutôt que l'intégration, nous avons favorisé l'interopérabilité en travaillant avec des fichiers de données faciles à produire dans le cadre d'un EIAH. Pour accorder DIANE et STAR, une fonction d'export des données a été mise en place dans l'environnement d'apprentissage. Un fichier d'extension .dataP est téléchargé et permet l'analyse dans STAR des données issues de DIANE. Nous n'excluons pas un usage de STAR de la part des professeurs des écoles puisqu'il est possible de leur faire télécharger une version de STAR dans laquelle les règles et le modèle sont préconstruits. Ces derniers n'ont alors plus qu'à charger le fichier .dataP et exécuter le modèle par son dernier onglet. Nous notons cependant qu'il faut être formé ou bien avoir une documentation pour comprendre la sortie du modèle (le taux de compression et la signification des différentes règles que constitue le diagnostic).

#### 9.2.2.2 Exemple: Formalisation d'un Knowledge Space

Notre formalisme basé sur les propriétés et les assemblages de règles est flexible, il est par exemple possible de créer un Knowledge Space de plusieurs manières. Un Knowledge Space est une représentation de l'organisation des compétences par la notion de prérequis.

**Cas ou le Knowledge Space est connu.**

Nous nous plaçons dans le cas où les problèmes sont étiquetés pour représenter les relations de prérequis, par exemple « NecessitePrerequisX, NecessitePrerequisY.... »

Dans ce cas, il suffit d'écrire la règle « absenceCapacitéX », qui a un « 1 » à l'intersection de :

- « échec » sur les propriétés réponses
- « NecessitePrerequisX » sur les propriétés problèmes.

Elle indique donc que si un apprenant répond à la règle « absenceCapacitéX » alors il risque d'échouer les problèmes qui ont pour prérequis cette capacité.

**Cas ou le Knowledge Space est recherché.**

Dans l'exemple précédent, les problèmes et réponses étaient déjà étiquetés en référence à un Knowledge Space connu.

Nous pouvons supposer, et c'est le principe de STAR, que certains doutes résident sur la construction de ce dernier. Une solution est donc de le représenter dans l'onglet Rules. Si une description suffisamment fine des propriétés problèmes existe, par exemple en incluant le niveau le plus précis de description, alors il est possible d'écrire des règles détaillées comme « absenceCapacitéX » prévoyant l'échec sur un ensemble précis de problèmes. En utilisant le même type de propriété et de règles, il est aussi possible d'écrire un modèle issu d'une Q-matrice.

#### 9.2.2.3 Flexibilité du formalisme

Les propriétés peuvent enrichir les réponses en tenant compte des caractéristiques du contexte. Les possibilités sont variées, nous donnons quelques exemples :

- La présence d'un élément facilitateur dans les problèmes
- L'utilisation de l'environnement en mode bac à sable
- Le temps (long ou court) de la résolution.

Par ces nouvelles propriétés, il est possible d'affiner les règles construites dans le modèle.

#### 9.2.2.4 Limites de l'aspect binaire des propriétés

STAR a été construit pour permettre à des modèles symboliques d'être exprimés et testés. De ce fait, la nature des données en entrée est préférentiellement symbolique. Toutefois, ce peut être une limitation, et une amélioration possible est de considérer qu'une propriété peut avoir une valeur numérique, ordinale ou catégorielle. Dans cette nouvelle approche, les règles ne s'exprimeraient plus comme la présence ou l'absence de propriétés, mais comme le résultat d'une fonction de filtre sur les propriétés. Ces filtres permettraient de conserver la forme symbolique du modèle même si une partie des données est numérique. Il serait par exemple possible d'indiquer qu'une règle s'applique, par exemple, seulement si 10>*âge*>12 ou *tempsResolution*>5s. « *âge* » et « *tempsRésolution* » seraient alors des propriétés auxquelles sont associées des valeurs numériques. Un tel développement alourdirait cependant l'approche de STAR qui se veut plutôt légère. Il faudrait des cas précis de règles qui ne peuvent s'exprimer que par ce niveau de précision supplémentaire pour justifier un tel effort.

#### 9.2.2.5 Limites du formalisme d'écriture de règles

Par son formalisme, STAR est paramétré par défaut pour écrire des contraintes. Il serait possible d'offrir un mode de construction de règles à l'exact opposé : une réponse est **autorisée** s'il existe au moins une règle qui **accepte** au moins une propriété de cette réponse sur ce problème. L'onglet Rules est plutôt flexible, il est donc possible, dans STAR, d'écrire une règle « précise » R1 indiquant que dans tel problème P, la réponse A seulement est attendue[44]. Par contre, la règle de cumul interne à STAR fonctionne par la négative. Si une autre règle R2 prédit la réponse A2 pour le problème P, alors la combinaison des règles R1 et R2 est l'ensemble vide. Il est envisageable de permettre un mode dans lequel les prédictions (et non plus les restrictions) sont cumulées. Une limitation similaire existe pour les propriétés sur les problèmes. Il n'est pas possible

[44] La tâche peut cependant être fastidieuse, car il faut indiquer des « block » sur toutes les lignes sauf celle souhaitée.

d'écrire une règle décrite comme valide seulement sur les problèmes qui ont **à la fois** la propriété P1, P2 et P3.

Nous avons imposé dans STAR une loi de composition interne adaptée à l'analyse par contraintes, c'est-à-dire un ensemble de règles qui, par cumul, restreignent l'espace des possibles. D'autres lois sont envisageables comme le « oui l'emporte » dans lequel des règles sont établies pour augmenter progressivement l'espace des possibles.

#### 9.2.2.6 Comparaison avec ASPM

Dans la partie théorique, nous avons mentionné le développement d'un algorithme d'optimisation de sélection de règles applicables à BUGGY : ASPM. STAR est un programme semblable à ASPM du point de vue de sa recherche du meilleur diagnostic, mais il diffère sur les règles qui le constituent. Dans ASPM elles font des prédictions uniques, ici il s'agit plutôt de prédictions multiples qui se combinent les unes aux autres. La distance à minimiser dans ASPM est la distance de Hamming (qui compte le nombre d'erreurs de prédiction). Dans notre approche, nous minimisons la taille du résumé des données formé par le diagnostic. Un avantage de STAR est son interface graphique. Il est difficile de déterminer les raisons pour lesquelles ASPM ne semble pas avoir été réutilisé dans d'autres projets, nous avons proposé des explications théoriques plus avant dans le manuscrit, mais son fonctionnement en ligne de commande a peut être rebuté les utilisateurs non experts.

Cependant STAR n'est pas optimisé pour simplifier la tâche de découverte de profil. En effet, l'algorithme employé est du type « force brute » : toutes les combinaisons possibles de règles sont recherchées et évaluées pour trouver la meilleure[45]. Cette approche peut être viable lorsque le nombre de règles est de l'ordre de la dizaine, mais lorsqu'il s'approche de la centaine, le nombre de diagnostics possibles peut alors devenir gigantesque ($2^{100} > 10^{30}$) et un algorithme d'optimisation semblable à ASPM pourrait être proposé.

[45] La recherche reste tout de même conditionnée au modèle qui décrit les exclusions inter-règles ce qui peut limiter l'espace de recherche de plusieurs ordres de grandeur.

## 9.3 Usage de STAR pour identifier les différences interindividuelles

## 9.4 De DIANE à STAR

Dans ce chapitre, nous montrons l'usage de STAR pour réaliser une analyse par contraintes dans le cadre de la résolution de PAEV. Nous avons réalisé une expérimentation à échelle réduite pour tester notre programme et travailler dans la continuité du modèle dynamique présenté plus tôt. Les passations ont été effectuées en utilisant DIANE. Un membre du laboratoire a développé un module de DIANE permettant d'exporter les données sous la forme requise par STAR. Ce module s'appuie sur le programme de diagnostic comportemental que nous avons développé et décrit en première partie de la thèse. Compte tenu des erreurs produites par le diagnostic, nous avons décidé de construire le module d'export sous une forme semi-automatique. À chaque réponse, DIANE donne une suggestion de diagnostic. Celle-ci se matérialise comme la présélection d'un diagnostic parmi un ensemble de réponses attendues. Elle est accompagnée de la valeur du compteur de fiabilité établi lors du diagnostic pour accompagner sa proposition d'une valeur de confiance. Chaque bloc est constitué d'une formule de calcul (sauf « pas de formule » et « ininterprétable » qui sont des cas particuliers) éventuellement accompagnée d'un ou plusieurs mots-clefs encodés dans un format spécifique à DIANE. La réponse présélectionnée par DIANE est encadrée, le nombre d'anomalies du diagnostic automatique est relevé. L'utilisateur clique sur un des blocs pour sélectionner le diagnostic comportemental qu'il pense approprié. Il le réalise problème par problème et élève par élève (Figure 44).

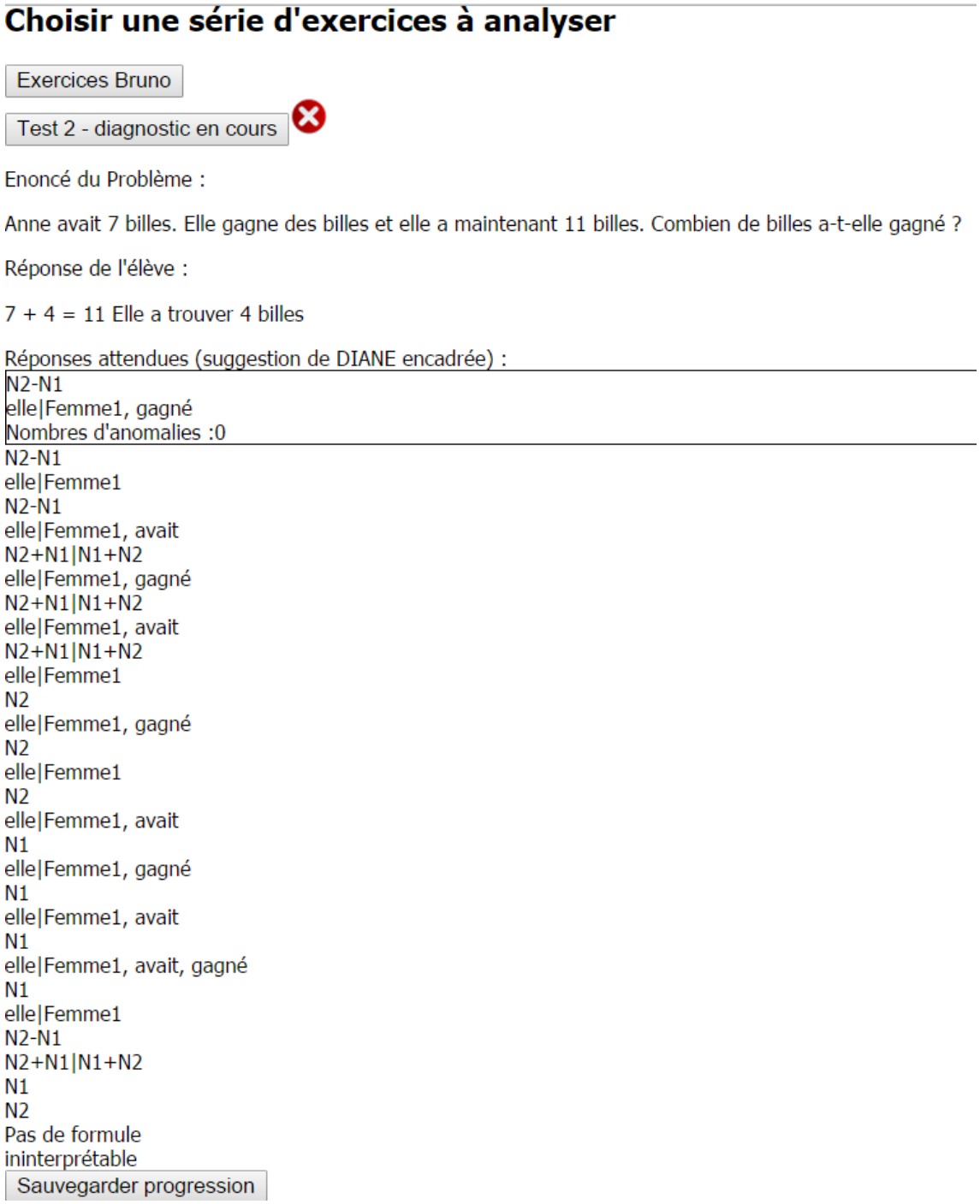



Figure 44. Écran de sélection de détermination du diagnostic. La suggestion du module de diagnostic comportemental est présélectionnée, accompagnée de la valeur du compteur de non-fiabilité. L'utilisateur clique sur le diagnostic comportemental qui lui semble le plus crédible.

Il est possible de sauvegarder la progression de l'activité de sélection, ce qui permet d'effectuer la tâche en plusieurs fois.

Lorsque l'utilisateur a fini d'identifier les réponses, l'archive « .dataP » est téléchargée.

### 9.4.1 Matériel

L'expérience concernait 46 élèves en classe élémentaire composés de niveaux variés : 20 CE1, 11 CE2, et 15 CM1. L'école se situe à Veynes, dans les Hautes-Alpes. L'expérience se déroulait en 2 fois 30 minutes. Les groupes étaient composés de 10 à 15 élèves, nous avons choisi de petits groupes en raison de la difficulté liée à l'encadrement d'élèves en primaire devant l'outil informatique afin d'éviter les phénomènes de copie ou de coopération, et de pouvoir rester disponible en cas de bugs informatiques. Malgré la simplicité de DIANE, une phase d'apprentissage était nécessaire pour s'assurer que tous les élèves maîtrisaient l'outil. Nous ne voulions pas faire une session d'apprentissage dans laquelle l'expérimentateur fait une démonstration, afin d'éviter la variabilité éventuelle dans la qualité de cette intervention. Nous avons produit **une vidéo interactive** permettant à chaque élève d'apprendre l'utilisation de DIANE de manière indépendante de l'expérimentateur et du groupe. Comme la capture d'écran en figure 45 en témoigne, la vidéo consistait à suivre un avatar montrant l'exemple de son inscription puis de l'utilisation de DIANE. Les élèves avaient simplement pour consigne de cliquer sur la flèche de droite lorsqu'ils avaient compris et la flèche de gauche dans le cas contraire. La flèche de gauche permettait de revenir en arrière et ainsi de rejouer l'audio et la vidéo.

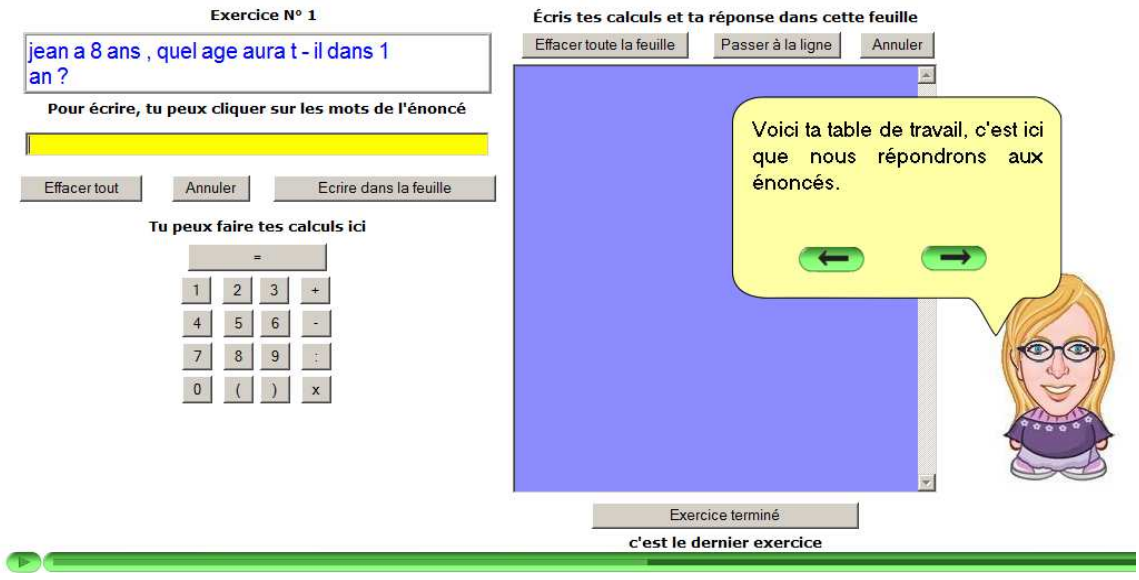



Figure 45. Capture d'écran de la vidéo interactive

Après avoir visionné la vidéo, les élèves avaient pour tâche de répondre à des problèmes similaires à ceux présentés en démonstration en utilisant l'interface de résolution (nommé « exercice simple » sur la Figure 45). Parmi les énoncés proposés figurent des énoncés dits « de remplissage ». Ces problèmes permettent d'éviter que l'élève repère

des régularités entre les problèmes et se base sur cette observation pour répondre. Ils sont au nombre de 3. Nous avons utilisé la possibilité que DIANE a d'organiser les énoncés en séries (Figure 46) pour pouvoir proposer 2 ordres différents pour les énoncés (et ainsi limiter les phénomènes de copie sur le voisin) et séparer les séries en une première comportant 6 exercices, dont 1 « de remplissage », et une deuxième comportant 9 exercices, dont 2 « de remplissage ».

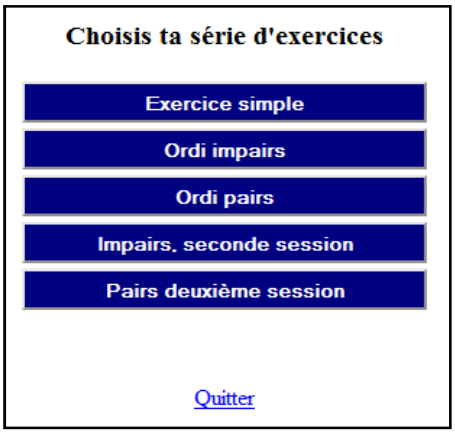



Figure 46. Organisation des séries dans DIANE

### 9.4.2 Problèmes et difficultés liés à l'expérimentation

Les problèmes ont été résolus en deux sessions. L'expérimentation n'a pas pu être complétée par tous les élèves (25/40) du fait de l'irruption d'activités non prévues au programme. Au moment de l'expérimentation, DIANE était différente de la version actuelle. Les diagnostics étaient réalisés au moment même ou les enfants résolvaient les problèmes. Or, dans des cas rares, un bug se produisait, ruinant la série d'exercices.

Nous supprimons de l'analyse les problèmes complexes à deux étapes avec question intermédiaire, d'une part parce que des pertes ont eu lieu sur ces problèmes, d'autre part parce que nous avons finalement choisi de simplifier les analyses. En effet nous n'avions pas d'hypothèse sur les implications d'une question intermédiaire dans les problèmes.

L'expérimentation n'est donc pas parfaite, toutefois les données restent analysables. Si la perte de données occasionnée est toujours dommageable, notre cadre d'analyse nous permet de rechercher des diagnostics même si certains problèmes n'ont pas été résolus. La conséquence est une diminution de la puissance des analyses, mais l'objectif consistant à mettre en évidence les différences individuelles n'est pas biaisé.

## 9.4.3 Modélisation dans STAR

Puisque l'objectif est d'exemplifier l'usage de STAR dans un cas concret et que nos données sont relativement faibles, un modèle de règle relativement simple est testé.

### 9.4.3.1 Modélisation du niveau de compétence

Dans chacune des catégories de problèmes représentées, leur difficulté est hiérarchisée. L'hypothèse souvent avancée est que le développement conceptuel de l'enfant ne lui permet de ne manipuler que certains schémas décrivant des relations mathématiques entre les quantités du problème.

**Note —** Dans cette partie, nous mettons en gras les contraintes et en italique les propriétés pour éviter au lecteur de possibles confusions.

Nous choisissons de construire un modèle de compétence hiérarchique basé sur cette échelle de difficulté. Ce modèle de compétence est relativement classique dans la littérature, il est proche de celui développé par Riley et Greeno (1988).

Nous avons choisi de mettre les deux problèmes complexes au niveau 3 de la hiérarchie, car bien qu'une planification soit nécessaire, les relations mathématiques ne sont pas plus compliquées dans ces problèmes. Comme nous le verrons plus tard, l'échec à ces seuls problèmes peut être cependant prédit par la contrainte de ne pas réutiliser des nombres.

| Type de Problème | Niveau 1 | Niveau 2 | Niveau 3 |
|---|---|---|---|
| Transformation | Lise avait 7 billes. Elle gagne 5 billes.<br>Combien de billes a-t-elle maintenant? | Anne avait 7 billes.<br>Elle gagne des billes et elle a maintenant 11 billes.<br>Combien de billes a-t-elle gagné ? | Simon gagne 6 billes et maintenant il a 13 billes.<br>Combien Simon avait-t-il de billes? |
| Comparaison | Jean a 8 billes<br>Carine a 5 billes de plus que Jean<br>Combien Carine a-t-elle de billes ? | Clara a 9 billes.<br>Tom a 14 billes.<br>Combien de billes Tom a-t-il de plus que Clara? | Léa a 11 billes<br>Elle a 5 billes de plus que Jérôme.<br>Combien Jérôme a-t-il de billes ? |
| Combinaison | Claire a 3 billes.<br>Pierre a 9 billes.<br>Combien de billes ont-ils ensemble ? | Paul a 4 billes.<br>Ensemble, Marie et Paul ont 11 billes.<br>Combien de billes Marie a-t-elle ? | |
| Complexe_1 | | | Zoé a 3 billes.<br>Sébastien a 5 billes de plus que Zoé.<br>Combien de billes ont-ils ensemble ? |
| Complexe_2 | | | Ben a 6 billes. Si il met ses billes avec Jessica, il y en a 14.<br>Combien Jessica a-t-elle de billes de plus que Ben ? |

Figure 47. Problèmes répartis selon leur type et leur niveau de difficulté

Les niveaux de compétences peuvent donc être décrits dans STAR sans difficulté de la manière suivante :

- Le **niveau 0** : contrainte interdisant la propriété *« bonne réponse »* sur l'ensemble des problèmes.
- Le **niveau 1** et **niveau 2** : contrainte interdisant la propriété *« bonne réponse »* sur un ensemble plus spécifique de problèmes.
- Le **niveau 3** : interdit la propriété *« absence de bonne réponse »*, ce qui permet de simplifier l'écriture de la contrainte.

#### 9.4.3.2 Règles alternatives au niveau de compétence

En opposition avec l'idée que l'échec aux problèmes provient d'une lacune dans le développement, nous avons eu l'occasion d'explorer deux hypothèses.

La première propose que les problèmes soient résolus par l'utilisation des mots-clefs, la deuxième qu'ils soient interprétés de manière parfois erronée. Ces deux hypothèses émettent des prédictions parfois en désaccord avec l'hypothèse d'échelle conceptuelle.

Dans cette partie, nous suggérons de voir ces hypothèses comme des traits individuels, possiblement en désaccord avec l'hypothèse développementale.

Deux autres règles sont donc écrites :

- **Accord mot-clef :** qui interdit l'absence de propriété « *Accord_mot_clef* » sur l'ensemble des problèmes. Cette propriété est donnée à toutes les réponses qui vont dans le même sens que le mot-clef principal du problème (ex. : gagner prédit une addition).
- **Accord interprétation** : qui interdit l'absence de la propriété « *reinterprétation_possible* » sur l'ensemble des problèmes. Cette propriété est associée à toute formule qui peut être comprise comme une interprétation bonne ou mauvaise du problème.

Enfin, nous avons proposé que les élèves se donnent des contraintes d'utilisation des nombres. Nous les décrivons aussi comme des règles possibles.

- **Pas de réutilisation** : interdit la propriété « *réutilisation_de_nombre* »
- **Usage de tous les nombres :** interdit l'absence de la propriété « *N1 utilisé* » et l'absence de la propriété « *N2 utilisé* ».

Seule une partie des propriétés dans DIANE est utilisée dans notre analyse. Nous en avons prévu d'autres, comme :

- *presence_verbal* : une verbalisation est donnée (e.g. Marie a 8 billes)
- *Absence Reponse* : l'enfant ne donne pas de réponse
- *Ininterprétable* : l'enfant ne donne pas une réponse compréhensible
- *Verbal accord_question* : la verbalisation suit les termes de la question
- *Verbal accord interprétation* : lorsque la verbalisation est en accord avec une réinterprétation erronée du problème.

Le Tableau 19 présente les réponses possibles au problème de transformation le plus difficile. Toutes les propriétés que nous avons pensé à établir dans DIANE sont représentées, y compris celles qui n'ont pas été utilisées dans le cadre de notre

modélisation. Dans ce tableau, les réponses se décomposent en 4 groupes. Le groupe vert (trois premières réponses) correspond aux bonnes réponses accompagnées d'une verbalisation ; le groupe rouge (10 réponses suivantes) des mauvaises réponses, aussi associées à des verbalisations ; le groupe bleu (4 réponses suivantes) comporte l'ensemble des réponses possibles sans verbalisation attendue ; le groupe gris (deux dernières réponses) représente les réponses « autres » (réponse vide, ininterprétable). Les propriétés utilisées dans la modélisation sont en gras.

Tableau 19. Tableau des réponses attendues à un problème et de leurs associations aux propriétés.

| Simon gagne 6 billes et maintenant il a 13 billes. Combien Simon avait-t-il de billes? | ***bonne_reponse*** | ***N1_utilisé*** | ***N2_utilisé*** | ***accord_mot_clef*** | ***reinterprétation_possible*** | *présence_verbal* | *verbal_accord_question* | *verbal_accord_reinterpretation* |
|---|---|---|---|---|---|---|---|---|
| **13-6=7 Simon avait 7 billes** | x | X | X | | X | X | X | X |
| 13-6=7 Simon a gagné 7 billes | X | X | X | | X | X | | X |
| 13-6=7 Simon a 7 billes | X | X | X | | X | X | | X |
| 13+6=19 Simon avait 19 billes | | X | X | X | X | X | X | X |
| 13+6=19 Simon a maintenant 19 billes | | X | X | X | X | X | | X |
| 13+6=19 Simon a gagné 19 billes | | X | X | X | X | X | | X |
| (pas de calcul) Simon a 6 billes | | X | | | X | X | | X |
| (pas de calcul) Simon a 6 billes | | X | | | X | X | | X |
| (pas de calcul) Simon avait 6 billes | | X | | | X | X | X | X |
| (pas de calcul) Simon avait gagné 6 billes | | X | | | X | X | X | X |
| (pas de calcul) simon avait 13 billes | | | X | | X | X | X | X |
| (pas de calcul) simon gagne 13 billes | | | X | | X | X | | X |
| (pas de calcul) simon a 13 billes | | | X | | X | X | | X |
| **6+13=19** | | X | X | X | X | | | |
| 13-6=7 | X | X | X | | X | | | |
| 6 | | X | | | X | | | |
| 13 | | | X | | X | | | |
| (réponse vide) | | | | | | | | |
| (ininterprétable) | | | | | | | | |

#### 9.4.3.3 Construction d'un modèle

Un modèle définit les assemblages possibles des règles créées.

En raison de notre hypothèse de travail, nous avons choisi d'opposer les trois modes de résolutions (conceptuelle, réinterprétation, mots-clefs) et de la complémenter par des heuristiques d'utilisation des nombres. La matrice de l'onglet Model a donc la forme représentée en Figure 48. Toutes les règles sont en relation d'exclusion sauf celles liées à l'utilisation des nombres. En comptant le profil « vide » (qui ne contient aucune règle), nous avons donc 28 profils possibles[46]. Le message compressé utilise donc pour encoder la réponse de chaque apprenant un préfixe de taille $Log_2(28)=4.8$ bits pour décrire le profil sélectionné de l'apprenant.

| | niveau 0 | niveau 1 | niveau 2 | niveau 3 | mots-clefs | reinterpretation | tous les nombres | pas de réutilisation de nombre |
|---|---|---|---|---|---|---|---|---|
| niveau 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| niveau 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| niveau 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| niveau 3 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| mots-clefs | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| reinterpretation | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| tous les nombres | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| pas de réutilisation de nombre | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Figure 48. Modèle décrit dans STAR dans le cadre d'une expérimentation.

#### 9.4.3.4 Sorties du modèle

Le modèle associe à chaque sujet :

[46] Détail du calcul : 6 règles de « mode de résolution » sont possibles : 4 de compétence, 2 alternatives, aucune règle sélectionnée, 7 en comptant l'absence de mode de résolution. De manière compatible aux autres, 2*2 possibilités d'utilisation des nombres sont possibles. Nous avons donc 7*4=28 profils possibles.

- Son diagnostic, c'est à dire l'ensemble des règles qui simulent au mieux son protocole.
- La compression liée au diagnostic. Elle peut être positive comme négative si le coût des paramètres du modèle n'est pas compensé par la diminution de la complexité des observations.
- Une p-value issue d'un test de permutation. Nous rappelons leurs significations dans ce cadre : une p-value à 0.2 signifie que lorsqu'une collection de réponses est tirée de l'échantillon, et qu'un diagnostic de l'élève « de Frankenstein » est effectué, la probabilité, obtenue par simulation, que son diagnostic ait un score meilleur que celui de l'élève courant, est de 0.2. Une p-value faible suggère donc que les caractéristiques individuelles du sujet ont été capturées. Nos calculs de p-values se font sur la base de 100 000 permutations.

## 9.4.4 Résultats

### 9.4.4.1 Pertinence des diagnostics

Tout comme les deux chapitres précédents, des ressources en ligne détaillant les analyses sont mises à disposition. Les liens vers ces ressources sont en annexe. Nous avons suggéré précédemment (cf. Tableau 18) que la mesure de la pertinence des diagnostics pouvait être établie par le calcul conjoint de deux mesures pour chaque enfant : la compression et la comparaison à des élèves « de Frankenstein » par un test de permutation. La compression apportée par le modèle est modérée (cf. Figure 49). Sa moyenne offerte par le modèle est négative (-0.77 bit) ce qui ne va pas dans le sens de notre modèle. En effet il n'est capable de compresser les données que pour une quantité réduite d'élèves. Comme nous l'avons précisé plus tôt, l'échec dans la compression peut être explicable par un ratio quantité de données sur le nombre de paramètres du modèle trop bas.

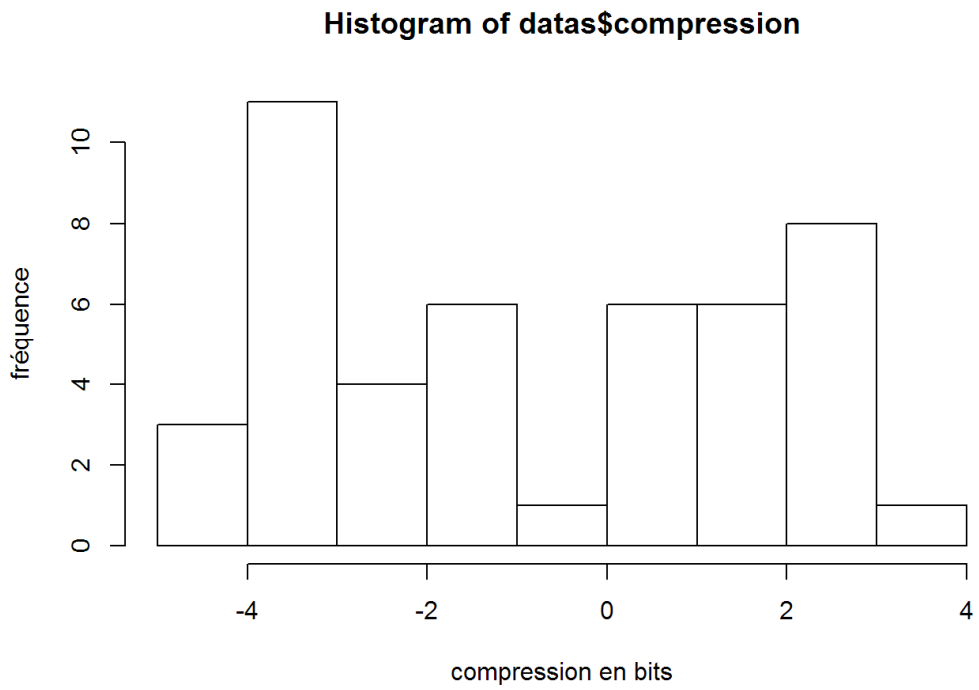



Figure 49. Compression en bits apportée par le modèle. Une compression positive indique que le modèle a pu compresser les données.

Cet échec relatif n'interdit pas non plus la possibilité de capturer des différences interindividuelles. Lorsque nous effectuons des tests de permutation pour quantifier cet aspect des diagnostics, nous obtenons des résultats plus satisfaisants pour une partie des diagnostics. La distribution des p-values est reportée en Figure 50. Cette distribution est encourageante dans la mesure où près de la moitié des élèves ont une p-value inférieure à 0.10.

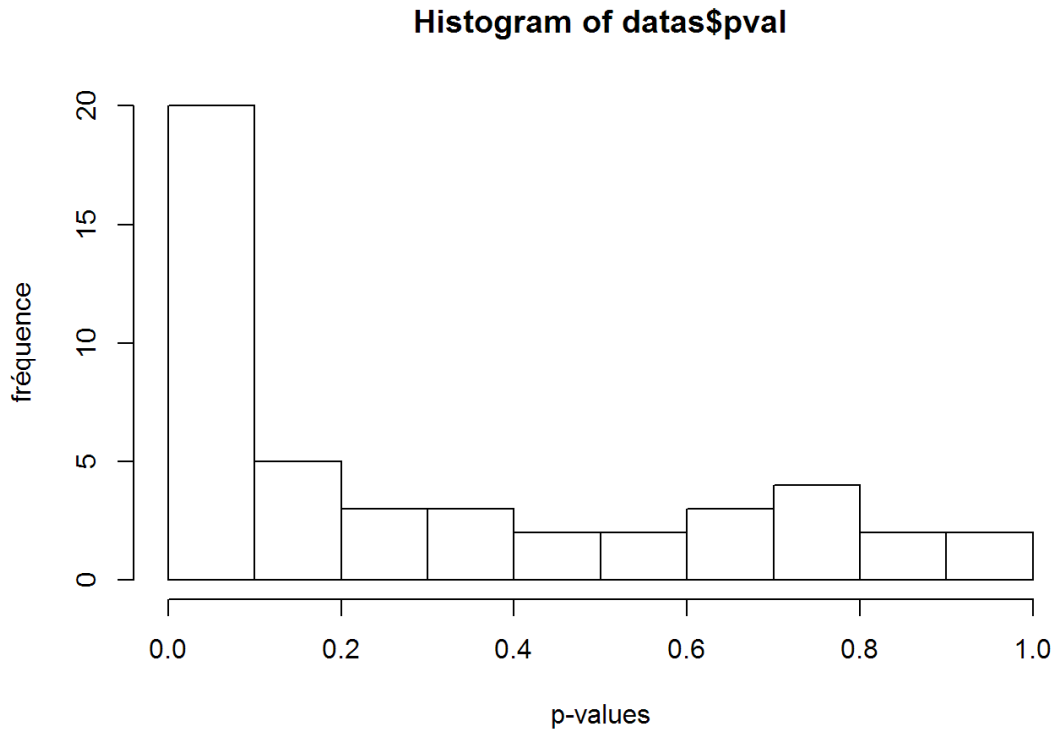



Figure 50. Histogramme décrivant la répartition des p-values

La répartition des p-values que nous avons tracée n'est pas suffisante pour tester si **d'un point de vue global**, le modèle a une pertinence pour capturer les différences interindividuelles. Cette réponse ne peut être obtenue que par la construction d'un test unique. La solution la plus esthétique que nous avons trouvée est d'agréger ces p-values par la méthode de Fisher[47]. Cette méthode permet d'obtenir une p-value globale en se basant sur la propriété connue selon laquelle **sous l'hypothèse nulle les p-values ont une distribution uniforme**. Il en suit que la somme des logarithmes des p-values

[47] Si tous les élèves étaient passés sur les mêmes problèmes, une solution plus directe pour construire un test global aurait été d'utiliser un test de permutation sur les moyennes dans les taux de compression pour répondre à la question statistique suivante :"quelle est la probabilité, en reconstituant 46 élèves de Frankenstein, d'obtenir une compression moyenne au moins aussi forte que la compression moyenne du groupe de départ ?". Cette méthode est inadéquate du fait que la compression n'est pas indépendante du nombre de problèmes répondus.

multipliée par (-2) suit une loi du Chi2 avec 2k degrés de liberté, k étant le nombre de p-values sommées.

$$X^2_{2k} \sim -2\sum_{i=1}^{k} \ln(p_i)$$

Le calcul du test unilatéral nous donne une p-value inférieure à $1.10^{-6}$ ce qui permet de nous assurer que le modèle ne capture pas du bruit.

#### 9.4.4.2 Contenu des diagnostics

Le Tableau 20 indique le nombre de fois où chaque règle est impliquée dans un diagnostic. Comme plusieurs règles peuvent faire partie d'un même diagnostic, le tableau somme à plus de 46.

Tableau 20. Nombre de fois où chaque règle a été diagnostiquée

| Niveau 0 | Niveau 1 | Niveau 2 | Niveau 3 | Accord mot-clef | Accord interprétation | Pas de réutilisation | Usage de tous les nombres |
|---|---|---|---|---|---|---|---|
| 2 | 7 | 3 | 27 | 2 | 4 | 37 | 7 |

Nous notons que les règles liées au niveau de compétence sont plus souvent diagnostiquées que celles décrivant une résolution par mots-clefs ou une par réinterprétation. Elles sont en relation exclusive. Les problèmes posés sont de difficulté croissante. Il n'est donc pas surprenant de voir que le niveau de compétence lève plus de complexité dans les protocoles que les deux autres modes de résolution.

Les contraintes portant sur l'usage des nombres ont souvent fait partie du diagnostic, notamment celle de non-réutilisation de nombres.

#### 9.4.4.3 Modèles alternatifs

Notre paradigme d'évaluation de diagnostic est basé sur un principe de sélection de modèles. Les règles Accord Mot-clef et Accord Interprétation sont relativement peu utilisées dans la formulation du diagnostic. Si nous les supprimons, deux évolutions du calcul de complexité allant dans le sens inverse l'une de l'autre vont avoir lieu :

- Moins de profils sont possibles (5x4=20). Donc, le coût pour indiquer celui sélectionné diminue, et ce, pour chaque élève. Il passe de $Log_2(28)$ à $Log_2(20)$ ce qui représente un gain de 0.49 bit.

- Ces règles ne faisant plus partie du diagnostic, pour les cas dans lesquelles elles étaient utilisées, la complexité pour décrire les élèves augmente, car un diagnostic moins adéquat leur est associé.

La question est donc de savoir si la perte de précision est compensée par la diminution de la complexité. Nous créons donc dans STAR une version du modèle dans laquelle deux « 0 » sont placés dans la diagonale de la matrice pour écarter ces règles et relançons les analyses.

Les résultats obtenus vont plutôt dans le sens de cette modification, même si l'apport n'est pas majeur. -0.32 plutôt que -0.77 bit pour l'ancien modèle. La significativité de cet apport peut être obtenue par un test de Wilcoxon pour échantillons appareillés ($V = 3$, p-value$<1.10^{-4}$ ).

#### 9.4.4.4 Discussion des résultats

Ces résultats semblent, au premier abord, en opposition avec ceux obtenus dans la deuxième partie des contributions sur les erreurs dans les problèmes complexes. Plusieurs éléments sont à prendre en compte pour la relativiser. Les problèmes ne représentent pas le même niveau de difficulté. En effet, la difficulté de ces problèmes les rendait très demandeurs en ressources cognitives. Par ailleurs, les analyses dans cette partie diffèrent fondamentalement, car seules les erreurs différentes de l'absence de réponse étaient analysées, ne représentant qu'une moitié des réponses obtenues. Ici, l'ambition était plus forte, nous avons cherché à mettre en évidence ces composantes de la résolution comme des traits individuels que l'apprenant conserverait sur tous les problèmes. Par ailleurs, l'activité de résolution par réinterprétation ou par mots-clefs a été mise en opposition avec un modèle de compétence. Ce dernier étant plus prédictif, les autres traits ont été écartés. Nous voyons ainsi une autre limite du formalisme de cumul de contrainte dans STAR. L'idéal aurait été de chercher un terrain partagé entre les différents modes de résolution du problème, par exemple en cherchant à diagnostiquer des règles mises en jeu uniquement en cas d'absence de solution. En l'occurrence, il aurait été souhaitable de proposer une contrainte d'interprétation ou de mot-clef qui laisse la priorité aux règles modélisant le niveau de l'apprenant. Ce mode d'écriture de règle n'est actuellement pas implémenté dans STAR.

## 9.5 Conclusion

Dans ce chapitre nous avons présenté une mesure basée sur le principe de minimum de complexité pour donner aux modèles non probabilistes les moyens d'être mesurés sans être dénaturés. Une telle mesure permet d'implémenter ou de guider les décisions dans la construction d'un modèle ou l'élaboration du diagnostic d'un élève.

STAR est issue de deux problématiques que nous considérons comme conjointes :

- La conception rapide du modèle général de diagnostic.
- L'évaluation de modèle de diagnostic sans passer par les approches psychométriques.

Le programme conçu calcule un critère adapté à une gamme de modèles cognitifs construits par des chercheurs en psychologie. STAR a pour but d'assister l'expert dans sa tentative de compréhension et de description des compétences et des conceptions déviantes ou erronées des sujets.

Nous avons proposé un calcul supplémentaire permettant de savoir si, outre la diminution de la complexité des données de l'apprenant, le diagnostic permettait de cerner les différences interindividuelles. En effet le but du diagnostic est de personnaliser les apprentissages, donc la capacité à cerner les différences entre les sujets est capitale. Cette mesure permet de soustraire de l'évaluation du modèle un certain nombre de biais évaluant à la hausse ses qualités prédictives par la modélisation de l'élève « moyen ». En effet, il est tout à fait possible d'avoir un modèle, qui du point de vue des métriques usuelles pour quantifier le pouvoir prédictif des modèles (Stamper, Koedinger, & McLaughlin, 2013 pour une revue), soit « bon » car prédisant les réponses les plus fréquentes, mais qu'il soit insensible ou presque aux différences interindividuelles. Pour tester cette sensibilité, il est possible, par la méthode que nous proposons, de quantifier à quel point le modèle est meilleur dans son diagnostic d'élèves réels que dans son diagnostic d'élèves reconstitués. Ce calcul peut être effectué quelque soit la statistique d'intérêt (pas seulement la réduction de complexité).

L'importance de cette contribution doit être relativisée du fait que randomiser les données des sujets pour étudier les qualités des modèles sur ces sujets reconstitués est une idée relativement simple et s'apparente à un test de permutation stratifié. Toutefois, notre analyse de la littérature ne nous a pas permis de trouver des approches similaires en modélisation de l'apprenant.

## 9.6 Discussion

STAR, en se limitant à des analyses basées sur le MDL brut, se prive des approches plus probabilistes, comme les réseaux bayésiens classiquement utilisés en Knowledge Tracing (Harrison & Roberts, 2012). STAR pourrait être éventuellement étendu dans la direction des modèles probabilistes et pourrait, par la même occasion, proposer des critères de sélection qui leur sont adaptés (MDL « moderne », AIC, BIC). Cependant, ne serait-ce pas un pas de trop vers la généricité ? En effet, il vise plutôt à la construction d'un espace d'écriture et d'évaluation pour des modèles symboliques. Nous avons précédemment insisté sur les limites du formalisme de STAR dans l'expression de ses règles. Il a été particulièrement pensé pour permettre une analyse du comportement par contraintes. Laisser la possibilité d'assembler les règles autrement que comme un cumul de contraintes (ou de se comporter différemment sur les propriétés problèmes et réponses) est possible dans un futur proche. Il faudrait construire une interface offrant plus d'options sans perdre l'utilisateur. L'essentiel est que le modèle conçu puisse produire une liste de profils possibles dont chacun produit des prédictions sur les données en entrée. Ainsi, si l'interface et les bibliothèques concernant les règles et le modèle peuvent être étendues, l'entrée (fichier dataP) et la sortie (calcul du MDL et établissement d'un test de permutation) peuvent rester inchangées.

L'idée directrice consistant à conserver les aspects non probabilistes des modèles a trouvé cependant ses limites. Sous certaines conditions (le cas où le profil choisi ne restreint pas suffisamment le champ des réponses), l'utilisation d'un bit « brut » pour coder le respect du profil est une simplification qu'il faut abandonner au profit d'une approche probabiliste (simplifiée).

STAR se base sur les propriétés associées aux réponses attendues et aux problèmes. Il est possible de l'utiliser en associant une propriété unique à chaque réponse possible et à chaque problème et ainsi abandonner la couche des propriétés. Cependant, il est alors difficile d'écrire des règles ayant du sens pour le concepteur de modèle cognitif. Nous pouvons aussi soulever un questionnement : est-il possible d'étiqueter des réponses efficacement sans avoir une idée du modèle qui va en rendre compte ? Dans notre modélisation, les propriétés telles qu'« accord mots-clefs » et « accord réinterprétation » montrent clairement une préparation à un modèle qui portera sur ces aspects. La construction d'un ensemble de réponses attendues peut souffrir du même problème. Elle est construite avec l'idée que chaque type de réponse a un sens différent.

Trois contre-arguments peuvent être fournis pour répondre à cette critique :

1. Les réponses attendues peuvent être justifiées par leurs fréquences d'apparition. Il est possible de voir un type de réponse apparaitre souvent, et le mettre dans le modèle justement pour essayer de l'expliquer par des règles. Certaines réponses attendues et certaines propriétés peuvent se passer d'un modèle pour exister. Par exemple : bonne réponse, mauvaise réponse, ininterprétable, pas de réponse... etc.

2. Dans le cadre d'une absolue méconnaissance du type de règles qui peuvent être construites, il est possible de simplifier le système de propriété en considérant que chaque problème et chaque réponse attendue ont une seule propriété : elle-même. Ainsi les règles dans STAR ne sont pas écrites **directement** comme une matrice dont les colonnes et les lignes sont les problèmes et les réponses attendues. Se débarrasser de la couche d'abstraction des propriétés permet d'éviter les impasses dans la construction des règles. Le prix à payer, cependant, est une perte du sens des règles écrites (et donc une perte de repères) d'où un travail rébarbatif dans l'écriture des règles.

3. Les règles ne sont pas le modèle. Sauf si l'indépendance des règles est assumée, il est nécessaire de définir leur organisation. Dans le cadre des conceptions erronées par exemple, certaines ne peuvent pas coexister, et d'autres sont en relation d'implication. S'il est possible d'avoir dès la conception une certaine assurance dans la forme des règles qui sont construites, leur organisation peut rester une zone d'indécision.

Une autre critique possible de STAR est que son utilisation excessive peut conduire à un phénomène de surajustement par le biais de la modification répétée d'un modèle. En effet, STAR est construit pour permettre une modélisation proactive avec l'évaluation. S'il est utilisé pour tester une large gamme de règles et de modèles, alors il est possible que certaines configurations apparaissent comme pertinentes « par chance ». Cependant, le surajustement causé par cette approche exploratoire est très faible comparée à une approche d'apprentissage automatique dans laquelle les « meilleures règles » et « meilleurs modèles » sont trouvés automatiquement. STAR n'implémente pas de tels algorithmes justement pour permettre au chercheur d'écrire et de tester des modèles qui font sens. L'activité de tester une grande quantité de modèles est coûteuse en temps et

contre-productive, dans la mesure où le modèle général obtenu a de faibles chances d'être généralisable. Le fait de pouvoir tester plusieurs modèles (ne serait-ce que deux ou trois) crée un biais de sélection. De ce fait, il peut être souhaitable d'analyser les données en deux phases, d'exploration et d'évaluation sur de nouvelles données pour obtenir des métriques fiables du modèle sélectionné.

Une approche souhaitable serait de comparer la nôtre à celle des modèles de compétences (espace de connaissances, Q-matrices). Les travaux originaux de Tatsuoka ont pour but de diagnostiquer des règles par cette méthode. Il serait par exemple pertinent de comparer, par le biais d'une simulation, la fiabilité d'un diagnostic lorsqu'il est obtenu par les méthodes psychométriques associées aux Q-matrices et lorsqu'il est obtenu par notre calcul basé sur le MDL. Dans le cadre spécifique de l'arithmétique en classe élémentaire, un espace de connaissance très détaillé a été mis en place dans l'environnement Merlin's Math Mill (Schoppek & Tulis, 2010). Ce modèle et cet environnement pourraient être un terrain idéal pour tester STAR.

Une objection possible à l'approche que nous avons développée porte sur l'aspect prédictif du diagnostic construit. Il pourrait être argué qu'attester du pouvoir explicatif du modèle n'est pas attester de son pouvoir prédictif. Deux éléments de réponses peuvent être donnés à cette critique. Le premier est de rappeler qu'une implémentation du MDL, lorsqu'il est appliqué aux modèles probabilistes, revient, à un terme près, à appliquer un critère purement construit sur le pouvoir prédictif du modèle : le BIC. Ensuite, il est facile de montrer qu'un modèle non prédictif, dans notre implémentation du MDL, ne permet pas de compresser les données. Un modèle non prédictif, par définition, ne permet pas d'apprendre des données, il donne des compressions équivalentes pour un enfant réel et pour un enfant fictif issu d'une permutation des données. Dans des termes logiques, la proposition « Non prédictif => Pas de compression » peut être renversée par « compression => prédictif ». La **quantification** de cette capacité prédictive n'est toutefois pas accessible. Nous notons que c'est aussi le cas pour des critères de sélection de modèles tel que l'AIC le BIC et la plupart des implémentations du MDL. Pour ces critères, la valeur absolue fournie n'a pas de sens dans l'absolu et permet simplement de comparer des modèles. Au contraire, la validation croisée permet de quantifier le pouvoir prédictif sur des données mises à l'écart. Des travaux pourraient être menés pour combiner les approches (MDL pour choisir le modèle et CV pour mesurer ses capacités prédictives). Ces travaux

permettraient aussi de préciser quantitativement la relation entre les capacités prédictives et explicatives du modèle.

# 10 CONCLUSIONS ET PERSPECTIVES

## 10.1 Résumé des contributions

Comme l'indiquent VanLehn (1988) et Wenger (1987), le diagnostic comportemental est un préalable au diagnostic épistémique, et celui-ci doit être riche pour ouvrir le champ des possibles. Pouvoir proposer et documenter une variété de problèmes est fondamental pour rendre un environnement d'apprentissage favorable à la modélisation cognitive. Nous avons conçu un outil auteur permettant de produire tout problème à structure additive. Lors de la création d'un problème, le concepteur est invité à le décrire de manière riche. Ce niveau d'exigence demandé, pouvant être limitant du point de vue du temps de création, permet de laisser ouvert son usage ultérieur. Tout modèle cognitif s'appuie sur des observations, et se doit d'être le plus riche et le plus fiable possible. Or, très peu d'environnements proposent une analyse des réponses ouvertes des apprenants. Nous avons construit un programme de diagnostic avec deux qualités qui le rendent unique dans la littérature : (1) générique, ce qui fait écho aux propriétés de l'outil auteur développé. Le module est capable de diagnostiquer tout problème à structure additive, sans lui donner la structure de ce problème ; (2) prudent. Il évite de faire des inférences trop risquées et fait remonter « ses doutes » en cas d'inférence modérément risquée. Nous avons pu évaluer la qualité du module et l'informativité de ses doutes, par la comparaison systématique avec le codeur humain sur une base de données, et par une expérimentation permettant de préciser notre connaissance sur la nature des désaccords. Le modèle de diagnostic a pu montrer de bonnes performances

avec 88 % d'accord avec le codeur humain sur la base de données considérée. Il a pu aussi montrer la légitimité de ses doutes. Lorsque son compteur de non fiabilité est à 0, les cas de désaccords avec l'humain sont rares (7 %). Lorsque les diagnostics humains et machines sont comparés, le diagnostic automatique semble l'emporter dans plus de la moitié des cas. Réciproquement, lorsque le compteur augmente, la quantité de désaccord avec l'humain est importante (37 %). Lorsque le diagnostic humain et machine sont comparés sur ces cas, l'humain emporte les votes dans la grande majorité des cas. En analysant la source des désaccords, nous avons cependant vu les limites d'un diagnostic comportemental qui ne prend pas en compte certains aspects comme les solutions expertes des problèmes analysés. La contribution associée à ce chapitre est aussi de nature méthodologique. Notre étude montre que rendre le diagnostic modulaire et autoévaluable est porteur d'informations. Il permet de mieux connaître les points qui peuvent être revus ou améliorés, mais aussi de préciser son usage possible dans un contexte de recherche en psychologie (utiliser l'humain pour corriger la machine ou le contraire).

Une fois les erreurs détectées, elles doivent pouvoir être comprises, c'est l'objet de notre conception de la modélisation cognitive en EIAH. De ce fait, la largeur de bande obtenue dans les travaux précédents n'est d'aucune utilité si elle n'est pas accompagnée d'une compréhension des différentes erreurs. Or, outre les modèles experts, la modélisation de la génération de réponse à un PAEV est peu présente dans la littérature des 30 dernières années. Nous avons conçu une modélisation, relativement modeste, de cette génération de réponse en nous basant sur l'hypothèse des réinterprétations des propositions du problème et sur l'hypothèse de l'utilisation des mots-clefs du problème pour la résolution de problème. Nous avons mis en place un programme informatique poussé et des méthodes d'analyses réfléchies pour tester ces hypothèses de la manière la plus appropriée possible. Des résultats substantiels ont été obtenus. Contrairement à nos hypothèses allant plutôt à l'encontre de l'idée d'une résolution par mots-clefs, nos résultats sont plutôt en faveur d'une coexistence des deux approches. De bons arguments existent donc pour les phénomènes de réinterprétations et de résolution par mots-clefs, même si notre étude a des angles morts qui mériteraient d'être étudiés en faisant varier le matériel. Cette partie est aussi porteuse d'une contribution méthodologique dans l'analyse de la qualité des prédictions de modèles génératifs déterministes. Une analyse construite sur la base des tests de permutation a permis

d'évaluer et de comparer des modèles en restant conforme au cadre d'analyse classique en psychologie cognitive passant par le calcul de p-values et de tailles d'effets.

L'ambition des EIAH est de pouvoir se centrer sur l'apprenant avec autant, si ce n'est plus, de pertinence qu'un humain. Ainsi, modéliser l'apprenant dans sa singularité est un enjeu et une problématique de taille. La conception de « modèle cognitif » est ambivalente et recouvre plusieurs approches. La notre se veut à rebours des deux tendances actuelles : (1) modéliser pour agir le plus vite possible lorsqu'une déviation du modèle expert est détectée (2) prendre le temps de former un diagnostic constitué se rapprochant d'une grille de compétences. Notre volonté est de modéliser le **comportement** dans sa régularité pour construire un **diagnostic**. Cette approche n'est pas nouvelle, mais restée en veille pour ce qui concerne les modèles symboliques. Nous avons modestement contribué à la réouverture de la porte aux modèles de règles en construisant une nouvelle mesure pour quantifier leur pertinence et un logiciel pour accélérer leurs développements. Cette mesure a pour racine le principe de minimum de complexité et permet de conserver la forme non probabiliste des modèles de règles. Dans cette conception, un modèle est pertinent s'il permet de réduire la complexité des données. Cette mesure permet de quantifier la pertinence d'un diagnostic en particulier, et, par accumulation, du système qui les produit. En complément de cette mesure, nous avons proposé une méthode basée sur la permutation des données des élèves pour évaluer à quel point le modèle cerne les différences interindividuelles. Sur la base de ces calculs, un logiciel (STAR) permettant d'écrire et de tester des modèles de règle a été construit. Nous avons alors illustré son fonctionnement dans une expérimentation à petite échelle visant à rechercher des différences interindividuelles dans le domaine de l'arithmétique. STAR travaille avec des données « étiquetées ». Les réponses et les problèmes sont augmentés par une liste de descripteurs et sont fournis par DIANE. STAR connecte ainsi les travaux sur DIANE (diagnostic comportemental et nouveau système de création de problèmes) et les travaux sur la modélisation des erreurs. Nous avons pu construire un modèle recherchant les régularités individuelles dans les phénomènes étudiés dans le chapitre sur la modélisation des erreurs. Les résultats fournis par STAR sont mixtes, ce qui est probablement dû à la faible quantité de données que nous avons utilisées pour l'analyse.

## 10.2 Généralisation des résultats

Nos travaux peuvent-ils être étendus à des problèmes différents ? Ce questionnement varie selon les types de problèmes envisagés. Traitons d'abord des problèmes à structures additives. Nos travaux sur la génération d'erreurs et sur le diagnostic comportemental portaient sur des problèmes avec une structure similaire. Outre l'aspect « étapes multiples » de ces problèmes, nous rappelons que deux stratégies de résolutions étaient disponibles (une ou trois étapes). La question de la généralisation de nos résultats doit donc être posée. D'un point de vue strict, il n'est pas possible d'affirmer que des problèmes dont la structure et la complexité diffèrent soient aussi faciles à diagnostiquer ou génèrent les mêmes comportements de résolutions mis en évidence.

En ce qui concerne la génération d'erreurs, les processus étudiés sont généraux et n'utilisent pas le fait, par exemple, que deux stratégies de résolutions sont possibles pour les résoudre. Le caractère essentiel de ces problèmes est leur complexité. La difficulté de résolution de ce type de problèmes agit comme un poids qui va permettre l'émergence de stratégies alternatives (comme la résolution par mots-clefs) ou de relâchements de contraintes (didactiques ou interprétatives). Pour des problèmes ne comportant que deux nombres, nos résultats auraient cependant plus de difficultés à être reproduits dans la mesure où l'espace de recherche engendré par la présence de deux nombres seulement est minuscule comparé à un espace constitué de 3 nombres. Si ces analyses étaient à reprendre sur un matériel nouveau, il serait au contraire pertinent d'employer des problèmes contenant quatre nombres. Une quantité pourrait éventuellement ne jouer aucun rôle dans le problème. Cette quantité, dans la théorie de la résolution par mots-clefs, pourrait cependant être employée dans les prédictions du modèle représentant cette stratégie, mais pas dans le modèle des réinterprétations. Ceci permettrait de distinguer avec plus de puissance les prédictions des deux modèles.

Du point de vue du diagnostic comportemental, ce type de problème pourrait aussi être intéressant à étudier. L'augmentation du nombre de quantités dans le problème pourrait provoquer des difficultés dans les opérations de désambigüisations de formules, car les sources possibles expliquant la présence d'un nombre inconnue sont statistiquement augmentées. En effet, nous avons établi que les modules cherchant à établir les sources des nombres utilisés manquaient d'efficacité, c'est pourquoi multiplier les sources peut poser un problème nouveau.

Il est aussi possible de se questionner sur les problèmes faisant intervenir d'autres opérations comme les problèmes de multiplication et de division. En ce qui concerne le diagnostic comportemental, des difficultés peuvent survenir. La nouvelle forme des opérations doit être prise en compte pour interpréter les formules et les éventuelles erreurs de calcul. La division pose peut-être plus de problèmes dans la mesure où les questions ou les étapes intermédiaires peuvent se baser sur le « reste » de l'opération. Autrement dit, un même calcul peut avoir deux productions qui peuvent être employées dans la suite des calculs. Il y a donc une ouverture du champ des possibles pour la division qui pourrait être problématique lorsque des démarches de désambigüisation doivent être réalisées.

Les problèmes arithmétiques ne sont pas les seuls problèmes scolaires basés sur un énoncé dont le texte décrit une situation. Toujours dans le domaine mathématique, certains problèmes d'algèbre contiennent un texte à interpréter. Les problèmes de physique sont aussi concernés, même si la description des problèmes est parfois mixte (en passant par exemple par des schémas). Pour ces problèmes, la démarche de notre modèle de diagnostic comportemental semble beaucoup moins applicable du fait de la variété d'opérations possibles dans le cas de la physique et de l'importance des explications. Dans les problèmes d'algèbre, la résolution n'est plus réalisée par une succession d'opérations, mais par la simplification progressive d'une (ou plusieurs) équation(s). La forme générale de notre diagnostic comportemental n'est plus tout à fait applicable.

## 10.3 Convergence des contributions

Cette thèse a été effectuée avec la ferme conviction que, sur le sujet de la modélisation du comportement, la psychologie cognitive et la recherche dans les environnements d'apprentissages peuvent travailler ensemble pour leur propre essor. L'état de l'art, notamment dans le domaine des PAEV a montré que cette dynamique de collaboration n'était pas complètement réalisée, autrement dit que la « fusion » ou le « mariage » des disciplines idéalisé il y a plusieurs décennies n'avaient pas eu lieu. Notre analyse de la littérature nous a permis de mieux saisir les différents obstacles pour ce rapprochement notamment les difficultés d'utilisation des modèles cognitifs dans le cadre de la production de diagnostics. La question est à la fois théorique (développement de modèles explicatifs) et technique (comment les exprimer, les mesurer, les employer dans un contexte interdisciplinaire). Chacune des contributions présentées a eu pour but

de proposer des pièces de ce puzzle dans le cadre de la résolution de PAEV. Les trois parties contributives répondent donc à des problématiques distinctes, mais convergentes, car elles cherchent à répondre à un problème de nature globale. Nous consacrons donc une partie pour préciser les synergies entre les différentes contributions présentées.

Les trois modèles construits ont une synergie particulière, ils peuvent outiller le chercheur dans ses recherches portant sur la résolution de PAEV variés. Les outils que nous avons développés sont relativement génériques, ce qui permet à d'autres théories, modèles ou problèmes d'être exprimées à l'aide de DIANE, de STAR et de notre programme de génération d'erreurs. Plusieurs scénarios sont imaginables. Le modèle de production d'erreurs pourrait être adapté pour tester d'autres hypothèses sur la résolution de problèmes complexes ou sur des problèmes mathématiques différents. Si ces hypothèses permettent de générer des réponses proches des observations dans une base de données alors le chercheur peut étudier, par l'emploi de STAR, la pertinence de ces hypothèses pour le diagnostic particulier des élèves. STAR lui permettrait, le cas échéant de tester plusieurs modèles cognitifs par le calcul intégré des taux de compression. DIANE pourrait jouer des rôles variés dans ce scénario. L'EIAH pourrait servir : (1) en amont pour, dans le cadre d'une analyse exploratoire, obtenir des données sur une large échelle sans chercher forcément à obtenir beaucoup de données pour un même élève ni contrôler finement les conditions d'expérimentation, (2) en position intermédiaire pour réaliser une expérimentation dans des classes et récupérer des données dans STAR au moyen de la passerelle existante. Dans une démarche exploratoire, le chercheur peut aussi modifier les propriétés sur les problèmes et sur les réponses attendues dans DIANE avant d'importer (ou de réimporter) les données dans STAR et affiner son modèle (3) en aval, pour intégrer dans DIANE un outil de diagnostic spécialisé basé sur les recherches menées. DIANE permet dans ce cas de collecter les données, mais aussi de réaliser un diagnostic comportemental automatique tout en repérant ses interprétations les moins fiables pour permettre à l'expérimentateur de corriger efficacement.

Notre thèse permet-elle l'idéal d'union entre la psychologie cognitive et les EIAH telle que nous l'avons esquissé en introduction ? À plus d'un titre, le chemin est encore long. Nous avons défendu l'idée que le développement de systèmes de diagnostic peut former un terrain d'entente. Or pour qu'un système de diagnostic puisse être porteur autant en psychologie que dans les environnements d'apprentissages, il faut qu'il puisse apporter

des informations fiables et utiles. Quant à l'utilité du diagnostic, c'est un point que nous avons laissé de côté dans notre thèse en présupposant qu'un diagnostic profond et fiable de l'apprenant avait de faibles chances d'être inutilisable d'un point de vue pédagogique. D'autres recherches doivent être menées, en didactique par exemple, pour étudier la pertinence des diagnostics de modèle symboliques. Certains outils, comme Pépite, utilisent le diagnostic pour la construction de groupes (Delozanne et al., 2008). Un autre intérêt du diagnostic serait d'outiller le professeur. Il constituerait, en d'autres termes, un éclairage cognitif des performances de sa classe ou de certains de ses élèves.

## 10.4 Perspectives

Le travail de génération d'erreurs a été établi dans un esprit de généricité. Il prend racine dans une approche générale basée sur le modèle des contraintes (Richard et al., 2009, 1993) qui est adaptée pour concilier des composantes hétérogènes dans la résolution de problème. Le modèle général de contraintes a agi ici comme un cadre d'analyse. Il ne peut être testé, mais peut inspirer l'implémentation d'un modèle, ses modes d'analyses et l'interprétation des résultats. Les modèles programmés permettent de générer des erreurs en fonction des paramètres donnés par l'utilisateur. Par cette modélisation, certains aspects de la résolution, comme les difficultés linguistiques et les stratégies par mots-clefs, ont pu être étudiés. D'autres facteurs ont été laissés de côté, notamment les effets de contenu ou la charge en mémoire de travail. Nous soulignons donc qu'il s'agissait moins de construire un modèle absolument général que de construire des modèles pour tester des hypothèses particulières sur la résolution de problème et d'évaluer leur potentiel pour expliquer certaines erreurs.

Si, comme nous l'avons souligné dans le paragraphe précédent, il est difficile d'avoir des certitudes sur la prolongation théorique de nos résultats sur d'autres problèmes, mais il est possible d'investiguer efficacement cette question par l'emploi de nos programmes. Ils sont développés à un haut niveau de généralité pour être utilisés sur d'autres données ou EIAH. En effet, le modèle de diagnostic comportemental ne prend en entrée que les nombres de l'énoncé et la réponse brute de l'enfant pour réaliser son diagnostic. Si d'autres bases de données sont mises à dispositions, l'établissement du diagnostic comportemental est pratiquement immédiat. Toutefois, sa validité ne peut-être mesurée qu'en présence d'un diagnostic comportemental concurrent pour lequel il est attribué un haut niveau de confiance. Vraisemblablement, le seul type de diagnostic répondant à cette exigence est le diagnostic humain réalisé par un expert.

La question de l'emploi du programme générant des erreurs sur d'autres problèmes est plus complexe. Le modèle des mots-clefs est relativement autonome et simple dans son implémentation, même si nous avons souligné que d'autres versions devraient être testées. Le modèle des réinterprétations est un peu plus complexe, car il cherche à résoudre le problème par l'emploi de réinterprétations préconstruites. L'humain, dans ce cadre, doit cependant décrire la structure mathématique du problème et les interprétations alternatives de chaque proposition pour pouvoir faire fonctionner le programme. Cette étape n'est pas excessivement longue, et le nombre de formulations mathématiques étant fini, il est possible de réemployer des blocs déjà construits dans notre implémentation. Une amélioration conséquente serait de construire un mécanisme générant ces réinterprétations. En d'autres termes, l'aspect génératif du modèle est limité s'il est comparé, par exemple, au programme Sierra qui produit des erreurs par l'emploi d'une théorie générale. Nous n'avons, de notre côté, pas d'hypothèses génératives sur la production des réinterprétations. Celles-ci ont été listées à la main sur la base de la littérature existante et sur un principe de similarité linguistique entre les différentes propositions.

Dans notre troisième partie, nous avons développé une méthode et un outil pour réactiver la place des modèles symboliques dans les diagnostics cognitifs. Dans la construction de cette méthode, nous avons explicitement critiqué le diagnostic basé sur la Rule Space Method issue de la psychométrie principalement parce qu'il ne tirait pas parti de la richesse des erreurs dans son diagnostic de règle erronée. Pour appuyer notre argumentaire, il serait intéressant, sur des données réelles ou simulées, de comparer la puissance statistique d'un diagnostic établi par notre méthode et d'un diagnostic établi par cette technique. Une possibilité serait alors d'étudier la propension de chaque méthode à établir des erreurs de diagnostic lorsque le niveau de bruit augmente (pour les données simulées) ou lorsque des données sont supprimées ou dégradées (pour des données réelles).

Un certain nombre de perspectives transversales peuvent être établies. Nos méthodes statistiques basées sur les tests de permutation pourraient être utilisées dans d'autres contextes. Il est difficile de se prononcer de manière définitive sur cette idée, mais notre étude de la littérature ne nous a pas permis de trouver de méthodes similaires pour la quantification de la pertinence de modèles cognitifs en utilisant une méthode basée sur les tests de permutations. Si les conditions s'y prêtent, nos méthodes pourraient être réemployées dans d'autres contextes. Dans notre troisième partie, l'utilisation d'une

permutation nous permet de savoir si notre modèle est sensible aux différences interindividuelles. Dans notre deuxième partie, cette même méthode nous permet de répondre de manière directe à la question « Quelle est la probabilité qu'un modèle faisant autant de prédictions (que notre modèle proposé) explique autant d'observations (que notre modèle proposé) ? ». Cette technique permet de contourner la difficulté posée par le caractère déterministe et binaire des prédictions tout en préservant la distribution réelle des observations. Il est plus difficile d'appliquer cette méthode à tous les modèles cognitifs déterministes, car il est nécessaire de disposer d'un espace légitime de prédictions possible dont les prédictions du modèle sont un sous-ensemble. Cette étape est délicate, dans notre étude, par exemple, la moindre fréquence des réponses selon leurs formes nous a conduits à introduire des tests stratifiés pour éviter que cette variabilité cause un biais dans nos analyses. Dans le cas de Sierra par exemple, la méthode n'est pas directement applicable, car il faut déterminer un espace de bugs ou d'observations possible qui fasse sens. Il est en effet difficile de construire et de justifier cet espace dans un domaine aussi ouvert que les soustractions posées. Si plusieurs modèles concurrents avaient amené des espaces de bugs différents alors il aurait été possible d'établir des analyses du même type en considérant que l'espace des bugs possibles est l'union de ces espaces et d'utiliser nos méthodes comparatives pour comparer les modèles entre eux. Cette méthode aurait pour avantage d'éviter d'utiliser plus d'une mesure pour comparer des modèles concurrents, mais elle ne peut faire sens que si l'espace de bugs issus de l'union des modèles est grand, ce qui ne semble pas être le cas actuellement du fait du manque de concurrents de Sierra.

Une autre perspective importante, transversale à nos trois contributions, serait de travailler avec des données plus riches que la « réponse finale » de l'apprenant. Par exemple, une réponse d'élève contient souvent des calculs effectués et parfois supprimés. Ces errements sont constitutifs à juste titre du protocole de résolution du sujet. Pour pouvoir les traiter de manière finalisée, il faut être capable conjointement de les détecter dans le cadre d'un EIAH (partie 1), les simuler dans le cadre d'une théorie explicative (partie 2), et les intégrer dans un diagnostic cognitif (partie 3) ce qui

représente un travail important, mais peut-être porteur. La richesse d'une réponse dépasse ces calculs abandonnés et une variété de pistes sont possibles allant de la dimension temporelle des actions dans la résolution (actuellement enregistrés dans DIANE, mais non traités) jusqu'aux mots qui la composent en passant par la forme des calculs[48]. Ces informations supplémentaires à condition d'implémenter leurs détections et de décider de leurs emplois peuvent affiner le diagnostic comportemental et cognitif du sujet en donnant plus de poids à certaines hypothèses selon leurs valeurs. Les mots utilisés pour justifier un calcul et la forme du calcul peuvent éventuellement être utilisés pour aider à la détermination de la présence ou non de réinterprétation. Les temps de résolutions et les errements peuvent aussi éclairer la nature des processus cognitifs engagés dans la résolution. Ces données peuvent aussi étayer des questions qui restent ouvertes dans notre deuxième partie. Par exemple, les erreurs qui s'expliquent seulement par le modèle par mot-clef sont-elles obtenues dans une résolution « rapide et directe » ou dans « une résolution lente et laborieuse » ? Répondre à cette question permettrait de savoir si les mots-clefs sont utilisés après un relâchement successif de contraintes ou comme une stratégie utilisée d'emblée. Dans le cadre de STAR ces données fournissent au programme d'autres dimensions qui peuvent être compressées à condition d'être capable de construire un modèle qui les explique et constituent donc, de manière légitime, une opportunité pour des modèles plus fins de montrer leurs valeurs.

[48] Dont la relation avec la représentation mentale du problème est bien établi (Brissiaud & Sander, 2010)

# 11 BIBLIOGRAPHIE

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, *7*(1), 39–59.

Adesina, A., Stone, R., Batmaz, F., & Jones, I. (2014). Touch Arithmetic: A process-based Computer-Aided Assessment approach for capture of problem solving steps in the context of elementary mathematics. *Computers & Education*, *78*, 333–343. http://doi.org/10.1016/j.compedu.2014.06.015

Akay, H., & Boz, N. (2010). The Effect of Problem Posing Oriented Analyses-II Course on the Attitudes toward Mathematics and Mathematics Self-Efficacy of Elementary Prospective Mathematics Teachers. *Australian Journal of Teacher Education*, *35*(1), 59–75.

Anderson, J. (1984). Cognitive psychology and intelligent tutoring. In *Proceedings of the Cognitive Science Society Conference* (pp. 37–43).

Anderson, J. R., & Schunn, C. (2000). Implications of the ACT-R learning theory: No magic bullets. *Advances in Instructional Psychology, Educational Design and Cognitive Science*, 1–33.

Atkins, J. (2005). The Association between the Use of Accelerated Math and Students' Math Achievement. Retrieved from http://dc.etsu.edu/etd/1028/

Balacheff, N. (1994). Didactique et intelligence artificielle. *Recherches En Didactique Des Mathématiques, 14*, 9–42.

Balacheff, N. (1995). Conception, connaissance et concept. In *Séminaire de l'équipe DidaTech, IMAG* (pp. 219–244). IMAG Grenoble. Retrieved from https://hal.archives-ouvertes.fr/hal-01072247/

Bartlett, F. C., & Burt, C. (1933). Remembering: A study in experimental and social psychology. *British Journal of Educational Psychology*, *3*(2), 187–192.

Baruk, S. (1985). *L'âge du capitaine: de l'erreur en mathématiques*. Éditions du Seuil.

Beck, J., Woolf, B. P., & Beal, C. R. (2000). ADVISOR: A machine learning architecture for intelligent tutor construction. *AAAI/IAAI, 2000*, 552–557.

Birch, M., & Beal, C. R. (2008). Problem Posing in AnimalWatch: An Interactive System for Student-Authored Content. In *FLAIRS Conference* (pp. 397–402). Retrieved from http://www.aaai.org/Papers/FLAIRS/2008/FLAIRS08-096.pdf

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16.

Bourgeois-Gironde, S., & Van Der Henst, J.-B. (2009). How to open the door to System 2: Debiasing the bat-and-ball problem. *Rational Animals, Irrational Humans*, 235–252.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*(3), 403–425.

Briars, D. J., & Larkin, J. H. (1984). An integrated model of skill in solving elementary word problems. *Cognition and Instruction*, *1*(3), 245–296.

Brissiaud, R. (1988). De l'âge du capitaine à l'âge du berger: Quel contrôle de la validité d'un énoncé de problème au CE2 ? *Revue Française de Pédagogie*, *82*(1), 23–31. http://doi.org/10.3406/rfp.1988.1457

Brissiaud, R., & Sander, E. (2010). Arithmetic word problem solving: a Situation Strategy First framework. *Developmental Science*, *13*(1), 92–107.

Brown, J. S., & Burton, R. R. (1978). Diagnostic Models for Procedural Bugs in Basic Mathematical Skills*. *Cognitive Science*, *2*(2), 155–192.

Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, *4*(4), 379–426.

Burton, R. (1982). DEBUGGY: Diagnosis of errors in basic mathematical skills. *Intelligent Tutoring Systems. Academic Press, London*.

Castro-Martínez, E., & Frías-Zorilla, A. (2013). Two-step arithmetic word problems. *Montana Mathematics Enthusiast*, *10*. Retrieved from http://www.math.umt.edu/TMME/vol10no1and2/15-Castro-Frias_pp379_406.pdf

Chaachoua, H., Nicaud, J. F., & Bittar, M. (2005). «Détermination automatique des théorèmes-en-acte des élèves en algèbre. Le cas des équations et inéquations de degré 1.». In *Actes de la conférence EIAH 2005* (pp. 33–45).

Chaachoua, H., Nicaud, J.-F., Bronner, A., & Bouhineau, D. (2004). Aplusix, a learning environment for algebra, actual use and benefits. In *ICME 10: 10th International Congress on Mathematical Education, July 4-11, 2004* (p. 8). Retrieved from http://hal.univ-grenoble-alpes.fr/hal-00190393/

Chaillet, V. (2014). *Les déterminants de la résolution de problèmes arithmétiques : Influence du caractère statique ou dynamique de l'énoncé sur le choix de la procédure et la nature des erreurs*. Retrieved from http://www.theses.fr/2014PA080035

Chaitin, G. J. (2005). Epistemology as information theory: From leibniz to omega. *arXiv Preprint math/0506552*. Retrieved from http://arxiv.org/abs/math/0506552

Chang, K.-E., Sung, Y.-T., & Lin, S.-F. (2006). Computer-assisted learning for mathematical problem solving. *Computers & Education*, *46*(2), 140–151. http://doi.org/10.1016/j.compedu.2004.08.002

Chang, K.-E., Wu, L.-J., Weng, S.-E., & Sung, Y.-T. (2012). Embedding game-based problem-solving phase into problem-posing system for mathematics learning. *Computers & Education*, *58*(2), 775–786. http://doi.org/10.1016/j.compedu.2011.10.002

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology: Section A*, *52*(2), 273–302.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Choi, D., & Ohlsson, S. (2010). Cognitive flexibility through learning from constraint violations. In *Proceedings of the Nineteenth Annual Conference on Behavior Representation in Modeling Simulation*. Retrieved from http://www.dongkyu.com/Resume_files/brims10_final.pdf

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, *40*(11), 4715–4729. http://doi.org/10.1016/j.eswa.2013.02.007

Cohen, P. R., Beal, C. R., & Adams, N. M. (2008). The Design, Deployment and Evaluation of the AnimalWatch Intelligent Tutoring System. (pp. 663–667). Retrieved from http://ebooks.iospress.nl/Download/Pdf/4455

Coquin-Viennot. (2000). Lecture d'énoncés de problèmes arithmétiques: effet d'une introduction thématique sur la construction de la représentation, pp. 41–58.

Coquin-Viennot, D. (2001). Problèmes arithmétiques verbaux à l'école: pourquoi les élèves ne répondent-ils pas à la question posée? *Enfance*, *53*(2), 181–196.

Corbett, A. T. (2001). Cognitive computer tutors: Solving the two-sigma problem. In *User Modeling 2001* (pp. 137–147). Springer.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 245–252). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=365111

Cummins, D. D. (1991). Children's interpretations of arithmetic word problems. *Cognition and Instruction*, *8*(3), 261–289.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*(4), 405–438.

d Baker, R. S., Pardos, Z. A., Gowda, S. M., Nooraei, B. B., & Heffernan, N. T. (2011). Ensembling predictions of student knowledge within intelligent tutoring systems. In *User Modeling, Adaption and Personalization* (pp. 13–24). Springer.

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H.-C. (2015). Word problems: a review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology*, *6*. http://doi.org/10.3389/fpsyg.2015.00348

De Corte, E., Verschaffel, L., & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology*, *77*(4), 460.

De Corte, E., Verschaffel, L., & Greer, B. (2000). Connecting mathematics problem solving to the real world. In *Proceedings of the International Conference on Mathematics Education into the 21st Century: Mathematics for living* (pp. 66–73). Retrieved from http://math.unipa.it/~grim/Jdecorte.PDF

De Corte, E., Verschaffel, L., & Pauwels, A. (1990). Influence of the semantic structure of word problems on second graders' eye movements. *Journal of Educational Psychology*, *82*(2), 359.

de la Torre, J. (2009). A Cognitive Diagnosis Model for Cognitively Based Multiple-Choice Options. *Applied Psychological Measurement*, *33*(3), 163–183. http://doi.org/10.1177/0146621608320523

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*(2), 269–273.

DeCarlo, L. T. (2011). On the Analysis of Fraction Subtraction Data: The DINA Model, Classification, Latent Class Sizes, and the Q-Matrix. *Applied Psychological Measurement*, *35*(1), 8–26. http://doi.org/10.1177/0146621610377081

Dehaene, S. (2001). Précis of the number sense. *Mind & Language*, *16*(1), 16–36.

Delozanne, É., Prévit, D., Grugeon, B., & Chenevotot, F. (2008). Automatic multi-criteria assessment of open-ended questions: a case study in school algebra. In *Intelligent Tutoring Systems* (pp. 101–110). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-69132-7_15

Delozanne, É., Prévit, D., Grugeon-Allys, B., & Chenevotot-Quentin, F. (2010). Vers un modèle de diagnostic de compétence. *Technique et Science Informatiques*, *29*(8–9), 899–938.

Depaepe, F., De Corte, E., & Verschaffel, L. (2015). Students' Non-realistic Mathematical Modeling as a Drawback of Teachers' Beliefs About and Approaches to Word Problem Solving. In *From beliefs to dynamic affect systems in mathematics education* (pp. 137–156). Springer.

Derry, S. J., Hawkes, L. W., & Tsai, C. (1987). A theory for remediating problem-solving skills of older children and adults. *Educational Psychologist*, *22*(1), 55–87.

Desmarais, M. C., Beheshti, B., & Naceur, R. (2012). Item to skills mapping: deriving a conjunctive q-matrix from data. In *Intelligent tutoring systems* (pp. 454–463). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-30950-2_58

Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based Q-Matrices. In *Artificial Intelligence in Education* (pp. 441–450). Springer.

Dessalles, J.-L. (2013). Algorithmic simplicity and relevance. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence* (pp. 119–130). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-44958-1_9

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31A Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models. In *Handbook of Statistics* (Vol. 26, pp. 979–1030). Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0169716106260310

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. *Cognitively Diagnostic Assessment*, 361–389.

Dijk, T. A. van, & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Dillenbourg, P., & Self, J. (1992). A framework for learner modelling. *Interactive Learning Environments*, *2*(2), 111–137.

Dubois, D., Nkambou, R., Quintal, J.-F., & Savard, F. (2010). Decision-making in cognitive tutoring systems. In *Advances in intelligent tutoring systems* (pp. 145–179). Springer.

Durand, C., & Vergnaud, G. (1976). Structures addictives et complexité psychogénétique. *Revue Française de Pédagogie*, *36*(1), 28–43. http://doi.org/10.3406/rfp.1976.1622

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54–75.

El-Kechaï, N., Delozanne, É., Prévit, D., Grugeon, B., & Chenevotot, F. (2011). Evaluating the performance of a Diagnosis System in School Algebra. In *Advances in Web-Based Learning-ICWL 2011* (pp. 263–272). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-25813-8_28

Elmadani, M., Mathews, M., & Mitrovic, A. (2012). Data-driven misconception discovery in constraint-based intelligent tutoring systems. Retrieved from http://ir.canterbury.ac.nz/handle/10092/7399

Fidler, F. (2002). The fifth edition of the APA Publication Manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, *62*(5), 749–770.

Fuchs, L. S., Compton, D. L., Fuchs, D., Powell, S. R., Schumacher, R. F., Hamlett, C. L., … Vukovic, R. K. (2012). Contributions of domain-general cognitive resources and different forms of arithmetic development to pre-algebraic knowledge. *Developmental Psychology*, *48*(5), 1315–1326. http://doi.org/10.1037/a0027475

Fuchs, L. S., Fuchs, D., Compton, D. L., Hamlett, C. L., & Wang, A. Y. (2015). Is Word-Problem Solving a Form of Text Comprehension? *Scientific Studies of Reading*, *19*(3), 204–223. http://doi.org/10.1080/10888438.2015.1005745

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., … Fletcher, J. M. (2006). The Cognitive Correlates of Third-Grade Skill in Arithmetic, Algorithmic Computation, and Arithmetic Word Problems. *Journal of Educational Psychology*, *98*(1), 29–43. http://doi.org/10.1037/0022-0663.98.1.29

Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing Mathematical Problem Solving Among Third-Grade Students With Schema-Based Instruction. *Journal of Educational Psychology*, *96*(4), 635–647. http://doi.org/10.1037/0022-0663.96.4.635

Fuson, K., Hudson, K., & Pillar, R. (1997). Phases of classroom mathematical problem-solving activity. *Employing Children's Natural Powers to Build Algebraic Reasoning in the Context of Elementary Mathematics*.

Gamo, S., Nogry, S., & Sander, E. (2014). RÉDUIRE LES EFFETS DE CONTENUS EN RÉSOLUTION DE PROBLÈMES POUR FAVORISER LA CONSTRUCTION D'UNE REPRÉSENTATION ALTERNATIVE. *Cahiers Des Sciences de l'Éducation–Université de Liège (aSPe)*, *36*, 35.

Gamo, S., Sander, E., & Richard, J. F. (2010). Transfer of strategy use by semantic recoding in arithmetic problem solving. *Learning and Instruction*, *20*(5), 400–410. http://doi.org/10.1016/j.learninstruc.2009.04.001

Gamo, S., Taabane, L., & Sander, E. (2011). Rôle de la nature des variables dans la résolution de problèmes additifs complexes. *Annee Psychologique*, *111*(4), 613.

Garcia, I., & Pacheco, C. (2013). A constructivist computational platform to support mathematics education in elementary school. *Computers & Education*, *66*, 25–39. http://doi.org/10.1016/j.compedu.2013.02.004

Gauvrit, N., Zenil, H., & Delahaye, J.-P. (2011). Assessing Cognitive Randomness: A Kolmogorov Complexity Approach. *arXiv Preprint arXiv:1106.3059*. Retrieved from http://www.researchgate.net/profile/Nicolas_Gauvrit/publication/51911767_Assessing_Cognitive_Randomness_A_Kolmogorov_Complexity_Approach/links/02e7e526c99fb3f662000000.pdf

Gerofsky, S. (1996). A linguistic and narrative view of word problems in mathematics education. *For the Learning of Mathematics*, 36–45.

Gerofsky, S. (1999). Genre analysis as a way of understanding pedagogy in mathematics education. *For the Learning of Mathematics*, 36–46.

Gerofsky, S. (2010). The impossibility of “real-life” word problems (according to Bakhtin, Lacan, Zizek and Baudrillard). *Discourse: Studies in the Cultural Politics of Education*, *31*(1), 61–73. http://doi.org/10.1080/01596300903465427

Gibbons, J. D., & Chakraborti, S. (2003). Nonparametric statistical inference fourth edition, revised and expanded. *STATISTICS TEXTBOOKS AND MONOGRAPHS*, *168*.

Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist*, *21*(1–2), 99–120.

Griffiths, T. L., & Tenenbaum, J. B. (2003). Probability, algorithmic complexity, and subjective randomness. In *Proceedings of the 25th annual conference of the cognitive science society* (pp. 480–485). Retrieved from http://www.its.caltech.edu/~theory/Gcompl.pdf

Grünwald, P. (1999). Viewing all models as “probabilistic.” In *Proceedings of the twelfth annual conference on Computational learning theory* (pp. 171–182). Retrieved from http://dl.acm.org/citation.cfm?id=307436

Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, Mass.: MIT Press.

Hakem, K., Chaillet, V., & Sander, E. (2011). DIANE, un EIAH fondé sur les effets de contenu pour les apprentissages arithmétiques: du diagnostic automatique à son interprétation. *EIAH 2011*, 301.

Hakem, K., Sander, E., & Labat, J. M. (2005). DIANE (Diagnostic Informatique sur l’Arithmétique au Niveau Elémentaire). Presented at the Environnements

Informatiques pour l'Apprentissage Humain (EIAH'2005), Montpellier. Retrieved from http://telearn.archives-ouvertes.fr/hal-00005704/

Hakem, K., Sander, E., Labat, J. M., & Richard, J. F. (2005). DIANE, a diagnosis system for arithmetical problem solving. *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, *125*, 258.

Harrison, B., & Roberts, D. L. (2012). A Review of Student Modeling Techniques in Intelligent Tutoring Systems. In *Eighth Artificial Intelligence and Interactive Digital Entertainment Conference*. Retrieved from http://www.aaai.org/ocs/index.php/AIIDE/AIIDE12/paper/download/5519/5778

Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology*, *84*(1), 76.

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, *87*(1), 18.

Hershkovitz, S., & Nesher, P. (1998). Tools to think with: Detecting different strategies in solving arithmetic word problems. *International Journal of Computers for Mathematical Learning*, *3*(3), 255–273.

Hirashima, T., Yokoyama, T., Okamoto, M., & Takeuchi, A. (2007). Learning by problem-posing as sentence-integration and experimental use. In *AIED* (Vol. 2007, pp. 254–261). Retrieved from http://books.google.com/books?hl=en&lr=&id=lAjvAgAAQBAJ&oi=fnd&pg=PA254&dq=%22Therefore,+as+a+future+work,+it+is+necessary+to+compare+the+learning%22+%22In+such+cases,+adequate+feedback+for+each+problem+i

s+required.%22+%22for+interactive+problem-posing+that+is+composed+of+(1)%22+&ots=ji1qY3RPTe&sig=uLxQx3RWnUE_No8SQ47i1dkyRAg

Huang, T.-H., Liu, Y.-C., & Chang, H.-C. (2012). Learning Achievement in Solving Word-Based Mathematical Questions through a Computer-Assisted Learning System. *Educational Technology & Society*, *15*(1), 248–259.

Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. *Child Development*, 84–90.

Inoue, N. (2005). The realistic reasons behind unrealistic solutions: the role of interpretive activity in word problem solving. *Learning and Instruction*, *15*(1), 69–83. http://doi.org/http://dx.doi.org/10.1016/j.learninstruc.2004.12.004

IREM de Grenoble. (n.d.). Quel est l'âge du capitaine? *Bulletin de l'APMEP*, (n° 323), 235–243.

Jaspers, M. W. M., & Van Lieshout, E. (1994). A CAI program for instructing text analysis and modelling of word problems to educable mentally retarded children. *Instructional Science*, *22*(2), 115–136.

Joséphine, T., Nkambou, R., & Bourdeau, J. (2006). A Framework to Specify a Cognitive Diagnosis Component in ILEs. *Journal of Interactive Learning Research*, *17*(3), 269–293.

Junker, B. W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. *Prepared for the National Research Council Committee on the Foundations of Assessment. Retrieved April*, *2*, 2001.

Käser, T., Baschera, G.-M., Kohn, J., Kucian, K., Richtmann, V., Grond, U., … von Aster, M. (2013). Design and evaluation of the computer-based training program

Calcularis for enhancing numerical cognition. *Frontiers in Psychology*, *4*. http://doi.org/10.3389/fpsyg.2013.00489

Katz, D., Baptista, J., Azen, S., & Pike, M. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, 469–474.

Kershaw, T. C., Flynn, C. K., & Gordon, L. T. (2013). Multiple paths to transfer and constraint relaxation in insight problem solving. *Thinking & Reasoning*, *19*(1), 96–136. http://doi.org/10.1080/13546783.2012.742852

Kieran, C. (1981). Concepts associated with the equality symbol. *Educational Studies in Mathematics*, *12*(3), 317–326.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, *92*(1), 109–129. http://doi.org/10.1037/0033-295X.92.1.109

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1534.

Koedinger, K. R., & MacLaren, B. A. (2002). Developing a pedagogical domain theory of early algebra problem solving. In *Carnegie Mellon University*. Citeseer.

Lajoie, S. P., & Derry, S. J. (2013). *Computers as cognitive tools*. Routledge. Retrieved from http://books.google.fr/books?hl=en&lr=&id=xGeasuCsN2MC&oi=fnd&pg=PR3&dq=Tutoring+Systems+and+Pedagogical+Theory:+Representational+Tools+

for+Understanding,+Planning,+and+Reflection+in+Problem+Solving&ots=O3lyA1rzRM&sig=AiMo443tQcCqXOhlR5EHj4CteD8

Lallé, S., Luengo, V., & Guin, N. (2012). Méthodologie d'assistance pour la comparaison de techniques de diagnostic des connaissances. *Intégration Technologique et Nouvelles Perspectives d'Usage*, 8.

Lallé, S., Mostow, J., Luengo, V., & Guin, N. (2013). Comparing Student Models in Different Formalisms by Predicting their Impact on Help Success. Retrieved from http://liris.cnrs.fr/Documents/Liris-6046.pdf

Langley, P., Wogulis, J., & Ohlsson, S. (1990). Rules and principles in cognitive diagnosis. *Diagnostic Monitoring of Skill and Knowledge Acquisition*, 217–250.

Le, N.-T., & Pinkwart, N. (2011). Enhancing the error diagnosis capability for constraint-based tutoring systems. In *Artificial Intelligence in Education* (pp. 500–503). Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-21869-9_82

LeBlanc, M. D., & Weber-Russell, S. (1996). Text integration and mathematical connections: a computer model of arithmetic word problem solving. *Cognitive Science*, *20*(3), 357–407.

Lee, M. D. (2005). An Efficient Method for the Minimum Description Length Evaluation of Deterministic Cognitive Models. In *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.138.4199&rep=rep1&type=pdf

Lee, M. D., & Cummins, T. D. (2004). Evidence accumulation in decision making: Unifying the “take the best” and the “rational” models. *Psychonomic Bulletin & Review*, *11*(2), 343–352.

Leh, J. M., & Jitendra, A. K. (2013). Effects of Computer-Mediated Versus Teacher-Mediated Instruction on the Mathematical Word Problem-Solving Performance of Third-Grade Students With Mathematical Difficulties. *Learning Disability Quarterly*, *36*(2), 68–79. http://doi.org/10.1177/0731948712461447

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: the role of cognitive models*. New York, NY: Cambridge University Press.

Lewis, A. B., & Mayer, R. E. (1987). Students’ miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology*, *79*(4), 363.

MacLaren, B., & Koedinger, K. (2002). When and why does mastery learning work: Instructional experiments with ACT-R “SimStudents.” In *Intelligent Tutoring Systems* (pp. 355–366). Springer.

Maier, N. R. (1930). Reasoning in humans. I. On direction. *Journal of Comparative Psychology*, *10*(2), 115.

Marshall, S. P. (1995). *Schemas in problem solving*. Cambridge: Cambridge University Press. Retrieved from http://0-dx.doi.org.oasis.unisa.ac.za/10.1017/CBO9780511527890

Martin, B., Labat, J.-M., & Sander, E. (2015). Synthèse des Environnements d'Apprentissage Abordant les Problèmes Arithmétiques en Classe Élémentaire. In *EIAH 2015*.

Martin, B., Sander, E., Labat, J.-M., & Richard, J.-F. (2013). STAR : Un outil permettant une intégration efficace de modèles cognitifs simples en EIAH. In *Actes de la conférence EIAH 2013* (pp. 315–326). Toulouse, France. Retrieved from http://hal.archives-ouvertes.fr/hal-00949170

Martin, B., Suraweera, P., & Mitrovic, A. (2007). Intelligent tutors for all: Constraint-based modeling methodology, systems and authoring. Retrieved from http://www.ir.canterbury.ac.nz/handle/10092/823

Mitrović, A. (1998). Experiences in Implementing Constraint-Based Modeling in SQL-Tutor. In B. P. Goettl, H. M. Halff, C. L. Redfield, & V. J. Shute (Eds.), *Intelligent Tutoring Systems* (Vol. 1452, pp. 414–423). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://link.springer.com/10.1007/3-540-68716-5_47

Mitrovic, A., Koedinger, K. R., & Martin, B. (2003). A comparative analysis of cognitive tutoring and constraint-based modeling. In *User Modeling 2003* (pp. 313–322). Springer. Retrieved from http://link.springer.com/chapter/10.1007/3-540-44963-9_42

Mitrovic, A., Martin, B., & Suraweera, P. (2007). Intelligent tutors for all: Constraint-based modeling methodology, systems and authoring. Retrieved from http://ir.canterbury.ac.nz/handle/10092/823

Mitrovic, A., Mayo, M., Suraweera, P., & Martin, B. (2001). Constraint-based tutors: a success story. In *Engineering of Intelligent Systems* (pp. 931–940). Springer. Retrieved from http://link.springer.com/chapter/10.1007/3-540-45517-5_103

Murray, T. (2003). An Overview of Intelligent Tutoring System Authoring Tools: Updated analysis of the state of the art. In *Authoring tools for advanced technology learning environments* (pp. 491–544). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-94-017-0819-7_17

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. *Handbook of Cognition*, 422–436.

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2014). Model Evaluation and Selection. *New Hanbook of Mathematical Psychology. Cambridge University Press, London*. Retrieved from http://faculty.psy.ohio-state.edu/myung/personal/MES-Dec2014.pdf

Nathan, M. J., Kintsch, W., & Young, E. (1990). *A theory of algebra word problem comprehension and its implications for unintelligent tutoring systems*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20.3906&rep=rep1&type=pdf

NCTM. (2000). *Principles and standards for school mathematics* (Vol. 1). National Council of Teachers of Mathematics.

Neches, R., Langley, P., & Klahr, D. (1987). *Learning, development, and production systems.* The MIT Press.

Nesher, P., & Hershkovitz, S. (1994). The role of schemes in two-step problems: Analysis and research findings. *Educational Studies in Mathematics, 26*(1), 1–23.

Nesher, P., & Katriel, T. (1977). A semantic analysis of addition and subtraction word problems in arithmetic. *Educational Studies in Mathematics*, *8*(3), 251–269.

Nesher, P., & Teubal, E. (1975). Verbal cues as an interfering factor in verbal problem solving. *Educational Studies in Mathematics*, *6*(1), 41–51.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. Retrieved from http://repository.cmu.edu/cgi/viewcontent.cgi?article=3032&context=compsci

Nicaud, J.-F., Chaachoua, H., & Bittar, M. (2006). Automatic calculation of students' conceptions in elementary algebra from Aplusix log files. In *Intelligent Tutoring Systems* (pp. 433–442). Retrieved from http://link.springer.com/chapter/10.1007/11774303_43

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). Cognitively diagnostic assessment. In *Based on the Conference on Alternative Diagnostic Assessment, U of Iowa Ctr for Conferences & Inst, May 1993.* Lawrence Erlbaum Associates, Inc.

Nogry, S., Guin, N., & Jean-Daubias, S. (2008). Ambre-add: An its to teach solving arithmetic word problems. *TECHNOLOGY INSTRUCTION COGNITION AND LEARNING*, *6*(1), 53.

Nortvedt, G. A. (2011). Coping strategies applied to comprehend multistep arithmetic word problems by students with above-average numeracy skills and below-average reading skills. *The Journal of Mathematical Behavior*, *30*(3), 255–269. http://doi.org/10.1016/j.jmathb.2011.04.003

Ohlsson, S. (1986). Some principles of intelligent tutoring. *Instructional Science*, *14*(3–4), 293–326.

Ohlsson, S. (1991). “ Viewpoint:” System Hacking Meets Learning Theory: Reflections on the Goals and Standards of Research in Artificial Intelligence and Education. *Journal of Interactive Learning Research*, *2*(3), 5.

Ohlsson, S. (1992). Constraint-based student modelling. *Journal of Interactive Learning Research*, *3*(4), 429.

Ohlsson, S. (1994). Constraint-based student modeling. In *Student modelling: the key to individualized knowledge-based instruction* (pp. 167–189). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-662-03037-0_7

Ohlsson, S. (2016). Constraint-Based Modeling: From Cognitive Theory to Computer Tutoring – and Back Again. *International Journal of Artificial Intelligence in Education*, *26*(1), 457–473. http://doi.org/10.1007/s40593-015-0075-7

Ohlsson, S., & Langley, P. (1988). *Psychological evaluation of path hypotheses in cognitive diagnosis*. Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4684-6350-7_3

Ohlsson, S., & Rees, E. (1990). *Adaptive search through constraint violations*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA225553

Öllinger, M., Jones, G., Faber, A. H., & Knoblich, G. (2013). Cognitive mechanisms of insight: The role of heuristics and representational change in solving the eight-coin problem. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 931–939. http://doi.org/10.1037/a0029194

Palm, T. (2006). Word problems as simulations of real-world situations: A proposed framework. *For the Learning of Mathematics*, 42–47.

Panaoura, A. (2012). Improving problem solving ability in mathematics by using a mathematical model: A computerized approach. *Computers in Human Behavior*, *28*(6), 2291–2297. http://doi.org/10.1016/j.chb.2012.06.036

Pastré, P. (1994). Évolution des compétences et formation: le cas de régleurs de presses à injecter. *Rapport de Recherche MESR–AGEFOS PME Bourgogne*, *91*.

Patrick, J., & Ahmed, A. (2014). Facilitating representation change in insight problems through training. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 532–543. http://doi.org/10.1037/a0034304

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Polya, G. (1945). *How to solve it; a new aspect of mathematical method* (Vol. xv). Princeton, NJ, US: Princeton University Press.

Posey, C. L., & Hawkes, L. W. (1996). Neural networks applied to knowledge acquisition in the student model. *Information Sciences*, *88*(1), 275–298.

Ragnemalm, E. L. (1995). Student diagnosis in practice; bridging a gap. *User Modeling and User-Adapted Interaction*, *5*(2), 93–116.

Reed, S. K. (1999). *Word problems research and curriculum reform*. Mahwah, N.J.: Lawrence Erlbaum Associates. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=19405

Reusser, K. (1990). Understanding Word Arithmetic Problems. Linguistic and Situational Factors. Retrieved from http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED326391

Reusser, K. (1993). Tutoring systems and pedagogical theory: Representational tools for understanding, planning, and reflection in problem solving. *Computers as Cognitive Tools, 1*, 143–177.

Reusser, K., & Stebler, R. (1997). Every word problem has a solution—the social rationality of mathematical modeling in schools. *Learning and Instruction, 7*(4), 309–327.

Reuter, T., Schnotz, W., & Rasch, R. (2014). Does representation matter? Teacher-provided tables and drawings as cognitive tools for solving non-routine word problems in primary school. In *Proceedings of the Frontiers in Mathematics and Science Education Research Conference*. Famagusta, North Cyprus. Retrieved from http://www.scimath.net/fiser2014/presentations/Timo%20Reuter.pdf

Richard, J. F., Pastré, P., & Parage, P. (2009). Analyse des stratégies de correction de défauts en plasturgie à l'aide d'un modèle de résolution de problème à base de contraintes. *Le Travail Humain, 72*(3), 267–292.

Richard, J. F., Poitrenaud, S., & Tijus, C. (1993). Problem-solving restructuration: Elimination of implicit constraints. *Cognitive Science, 17*(4), 497–529.

Richard, J. F., & Sander, E. (2000). Activités d'interprétation et de recherche de solutions dans la résolution de problèmes. *J.–N. Foulin & C. Ponce (Eds.), Lire, Écrire, Compter, Apprendre: Les Apports de La Psychologie Des Apprentissages, Editions Du CRDP de Bordeaux*, 91–102.

Riley, M. S., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction, 5*(1), 49–101.

Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, 1080–1100.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. T. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249–255.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358.

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: A MDL-Based Cognitive Model of Inductive Learning. *Cognitive Science*, *35*(7), 1352–1389. http://doi.org/10.1111/j.1551-6709.2011.01188.x

Roe, A., & Taube, K. (2004). How Can Reading Abilities Explain Differences in Maths Performances? *Northern Lights on PISA 2003: A Reflection from the Nordic Countries*.

Ruokamo, H., & Pohjolainen, S. (1998). Pedagogical principles for evaluation of hypermedia-based learning environments in mathematics. *Journal of Universal Computer Science*, *4*(3), 292–307.

Rupp, A. A., & Templin, J. L. (2008). Unique Characteristics of Diagnostic Classification Models: A Comprehensive Review of the Current State-of-the-Art. *Measurement: Interdisciplinary Research & Perspective*, *6*(4), 219–262. http://doi.org/10.1080/15366360802490866

Sander, E. (2001). Solving arithmetic operations: a semantic approach. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 915–920). Edinburgh Scotland.

Sander, E. (2007). Manipuler l'habillage d'un problème pour évaluer les apprentissages. *Bulletin de Psychologie*, 119–124.

Sander, E., & Richard, J.-F. (2005). Analogy and transfer: encoding the problem at the right level of abstraction. Retrieved from http://www.psych.unito.it/csc/cogsci05/frame/poster/2/f225-sander.pdf

Sarrazy, B. (2002). Effects of variability of teaching on responsiveness to the didactic contract in arithmetic problem-solving among pupils of 9–10 years. *European Journal of Psychology of Education*, *17*(4), 321–341.

Schoppek, W., & Landgraf, A. (2011). Can a multidimensional hierarchy of skills generate data conforming to the Rasch model? A comparison of methods. *Psychological Test and Assessment Modeling*, *53*, 3–34.

Schoppek, W., & Tulis, M. (2010). Enhancing Arithmetic and Word-Problem Solving Skills Efficiently by Individualized Computer-Assisted Practice. *The Journal of Educational Research*, *103*(4), 239–252. http://doi.org/10.1080/00220670903382962

Schumacher, R. F., & Fuchs, L. S. (2012). Does understanding relational terminology mediate effects of intervention on compare word problems? *Journal of Experimental Child Psychology*, *111*(4), 607–628. http://doi.org/10.1016/j.jecp.2011.12.001

Self, J. (1993). Model-based cognitive diagnosis. *User Modeling and User-Adapted Interaction*, *3*(1), 89–106.

Self, J. A. (1990). Bypassing the intractable problem of student modelling. *Intelligent Tutoring Systems: At the Crossroads of Artificial Intelligence and Education*, 107–123.

Self, J. A. (1994). Formal approaches to student modelling. *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, *125*, 295–295.

Seo, Y.-J., & Woo, H. (2010). The identification, implementation, and evaluation of critical user interface design features of computer-assisted instruction programs in mathematics for students with learning disabilities. *Computers & Education*, *55*(1), 363–377. http://doi.org/10.1016/j.compedu.2010.02.002

Sheriff, K. A., & Boon, R. T. (2014). Effects of computer-based graphic organizers to solve one-step word problems for middle school students with mild intellectual disability: A preliminary study. *Research in Developmental Disabilities*, *35*(8), 1828–1837. http://doi.org/10.1016/j.ridd.2014.03.023

Simon, H. A. (1990). Prediction and prescription in systems modeling. *Operations Research*, *38*(1), 7–14.

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, *26*(2), 145–159. http://doi.org/10.1037/h0030806

Simon, H. A., Polk, T., & Vanlehn, K. (1995). *Analysis of Symbolic Parameter Models.* DTIC Document.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Cambridge, Mass, USA*, 99–113.

Snow, R. E., & Lohman, D. F. (1989). *Implications of cognitive psychology for educational measurement.* American Council on Education.

Stamper, J. C., Koedinger, K. R., & McLaughlin, E. A. (2013). A Comparison of Model Selection Metrics in DataShop. In *EDM* (pp. 284–287). Retrieved from http://educationaldatamining.org/EDM2013/proceedings/paper_75.pdf

Steffe, L. P., & Olive, J. (2002). Design and use of computer tools for interactive mathematical activity (TIMA). *Journal of Educational Computing Research*, *27*(1), 55–76.

Swanson, H. L., Cooney, J. B., & Brock, S. (1993). The Influence of Working Memory and Classification Ability on Children′s Word Problem Solution. *Journal of Experimental Child Psychology*, *55*(3), 374–395. http://doi.org/http://dx.doi.org/10.1006/jecp.1993.1021

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational and Behavioral Statistics*, *10*(1), 55–73.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. *Diagnostic Monitoring of Skill and Knowledge Acquisition*, 453–488.

Tatsuoka, K. K. (2009). *Cognitive assessment: an introduction to the rule space method*. New York: Routledge.

Thevenot, C., & Oakhill, J. (2008). A generalization of the representational change theory from insight to non-insight problems: The case of arithmetic word problems. *Acta Psychologica*, *129*(3), 315–324. http://doi.org/10.1016/j.actpsy.2008.08.008

Tsuei, M.-P. (2007). A web-based curriculum-based measurement system for class-wide ongoing assessment: ECBM system for ongoing assessment. *Journal of*

*Computer Assisted Learning*, *24*(1), 47–60. http://doi.org/10.1111/j.1365-2729.2007.00242.x

van der Schoot, M., Bakker Arkema, A. H., Horsley, T. M., & van Lieshout, E. C. D. M. (2009). The consistency effect depends on markedness in less successful but not successful problem solvers: An eye movement study in primary school children. *Contemporary Educational Psychology*, *34*(1), 58–66. http://doi.org/10.1016/j.cedpsych.2008.07.002

VanLehn, K. (1988). Student modeling. *Foundations of Intelligent Tutoring Systems*, *55*, 78.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. The MIT Press. Retrieved from http://books.google.fr/books?hl=en&lr=&id=owY9seMJ4RwC&oi=fnd&pg=PR9&dq=Mind+Bugs++The+Origins+of+Procedural+Misconceptions++Learning++Development++and+Conceptual+Change&ots=vN9isplWdE&sig=XdnJyFxmy3wceZw0Ol31mjiJzlg

Vergnaud, G. (1982). A classification of cognitive tasks and operations of thought involved in addition and subtraction problems. *Addition and Subtraction: A Cognitive Perspective*, 39–59.

Verschaffel, L. (1994). Using retelling data to study elementary school children's representations and solutions of compare problems. *Journal for Research in Mathematics Education*, 141–165.

Verschaffel, L. (2000). *Making sense of word problems*. Lisse [Netherlands]; Exton, PA: Swets & Zeitlinger Publishers.

Verschaffel, L., Corte, E. D., & Lasure, S. (1994). Realistic considerations in mathematical modeling of school arithmetic word problems. *Learning and Instruction*, *4*(4), 273–294. http://doi.org/http://dx.doi.org/10.1016/0959-4752(94)90002-7

Verschaffel, L., De Corte, E., & Pauwels, A. (1992). Solving compare problems: An eye movement test of Lewis and Mayer's consistency hypothesis. *Journal of Educational Psychology*, *84*(1), 85.

Verschaffel, L., Dooren, W., Greer, B., & Mukhopadhyay, S. (2010). Reconceptualising Word Problems as Exercises in Mathematical Modelling. *Journal Für Mathematik-Didaktik*, *31*(1), 9–29. http://doi.org/10.1007/s13138-010-0007-x

Vicente, S., Orrantia, J., & Verschaffel, L. (2007). Influence of situational and conceptual rewording on word problem solving. *British Journal of Educational Psychology*, *77*(4), 829–848. http://doi.org/10.1348/000709907X178200

Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J.-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, *28*(4), 409–426.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children*, 34–41.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge*. Los Altos, Calif: Morgan Kaufmann Publishers.

Willis, G. B., & Fuson, K. C. (1988). Teaching children to use schematic drawings to solve addition and subtraction word problems. *Journal of Educational Psychology*, *80*(2), 192.

Wilson, A. J., Dehaene, S., Pinel, P., Revkin, S. K., Cohen, L., & Cohen, D. (2006). Principles underlying the design of “The Number Race”, an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions*, *2*(1), 19.

Woolf, B. P. (2009). *Building intelligent interactive tutors: student-centered strategies for revolutionizing e-learning*. Amsterdam: Elsevier [u.a.].

Yudelson, M., Pavlik Jr, P. I., & Koedinger, K. R. (2011). User Modeling–A Notoriously Black Art. In *User Modeling, Adaption and Personalization* (pp. 317–328). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-22362-4_27

# 12 ANNEXES



CODE ET ANALYSES STATISTIQUES MIS EN LIGNE ..... 276

PROBLÈMES ..... 277

PROBLÈMES — VERSION 2 ..... 280

CRÉATION D'UN NOUVEAU MODE DE CRÉATION DE PROBLÈME DANS DIANE ..... 283

CONSIGNES DANS L'EXPÉRIMENTATION ÉVALUANT LE DIAGNOSTIC COMPORTEMENTAL ..... 290

# CODE ET ANALYSES STATISTIQUES MIS EN LIGNE

Dans une démarche d'ouverture er de reproductibilité, nous avons jugé important de mettre en ligne nos données, programmes et analyses, même si ces ressources sont encore imparfaites en terme de documentation. Nous présentons du code, mais aussi des "notebooks" (IPython pour le langage python et knitr son équivalent pour R). Ceux-ci permettent de produire du code ou des analyses statistiques tout en facilitant leur présentation. Tous les fichiers présentés sont placés sur Github qui est un outil en ligne populaire pour favoriser le partage et la collaboration. Les résultats des analyses statistiques sont sous forme HTML, **pour les visualiser, nous recommandons d'utiliser le site https://htmlpreview.github.io/** en lui fournissant l'adresse du fichier HTML souhaité.

**Diagnostic comportemental :**

- Code source de DIANE : https://github.com/brumar/DIANE
- Code source du diagnostic comportemental : https://github.com/brumar/Behavioral-Diagnosis-of-Word-Problem-Solving
- Analyse statistique des performances du diagnostic comportemental (expérimentation comprise) : https://github.com/brumar/Behavioral-Diagnosis-of-Word-Problem-Solving/tree/master/R_Analysis

**Modélisation cognitive :**

- Code source du programme de génération d'erreur : https://github.com/brumar/WPsolving (le fichier *main.py*, en particulier qui paramètre la simulation). Au sein de ce dépôt, des notebooks (fichiers d'extension *ipynb*) sont présentés. L'un d'eux, *MultiStep Word Problems investigation.ipynb*, est celui qui est présenté dans le chapitre 9.
- Analyses statistiques sur l'ensemble des modèles et leurs comparaisons (y compris la contrainte de soustraction analysée de manière post-hoc) : https://github.com/brumar/WPsolving/blob/master/R_Analysis

**Diagnostic épistémique :**

- Code source de STAR : https://github.com/brumar/STAR
- Analyses statistiques liées à l'expérimentation : https://github.com/brumar/STAR/tree/master/MC_R_analysis

# PROBLEMES

**Tc1t**

Pendant la récréation, Lucas gagne 7 billes. Après la récréation, Lucas a 16 billes. Avant la récréation, Simon avait autant de billes que Lucas. Pendant la récréation, Simon gagne 3 billes de moins que Lucas. Combien Simon a-t-il de billes après la récréation ?

**Tc1p**

Pendant la récréation, Lucas gagne 7 billes. Après la récréation, Lucas a 16 billes. Avant la récréation, Simon avait autant de billes que Lucas. Pendant la récréation, Simon gagne des billes, et après la récréation, il a 3 billes de moins que Lucas. Combien Simon a-t-il gagné de billes pendant la récréation ?

**Tc2t**

Cette année, Théo a été pesé par le pédiatre. Théo a pris 5 kilos depuis le début de l'année. Théo pèse maintenant 14 kilos. Au début de l'année, Nicolas pesait le même poids que Théo. Nicolas a pris 2 kilos de moins que Théo cette année. Combien Nicolas pèse-t-il maintenant ?

**Tc2p**

Cette année, Théo a été pesé par le pédiatre. Théo a pris 5 kilos depuis le début de l'année. Théo pèse maintenant 14 kilos. Au début de l'année, Nicolas pesait le même poids que Théo. Maintenant, Nicolas pèse 2 kilos de moins que Théo. Combien de kilos Nicolas a-t-il pris cette année ?

**Tc3t**

En janvier, 7 enfants se sont inscrits à la chorale. Après janvier, il y a 16 enfants à la chorale. Avant janvier, il y avait autant d'enfants inscrits au football qu'à la chorale. En janvier, il y a eu 2 inscriptions de moins au football qu'à la chorale. Combien y a-t-il d'enfants au football après janvier ?

**Tc3p**

En janvier, 7 enfants se sont inscrits à la chorale. Après janvier, il y a 16 enfants à la chorale. Avant janvier, il y avait autant d'enfants inscrits au football qu'à la chorale. En janvier, il y a eu de nouvelles inscriptions au football. Après janvier, il y a 2 enfants de moins au football qu'à la chorale. Combien d'enfants se sont inscrits au football en janvier ?

**Tc4t**

Au supermarché, le kilo de poisson a augmenté de 5 euros cette année. Un kilo de poisson coûte maintenant 12 euros. Au début de l'année, le kilo de viande coûtait le même prix que le kilo de poisson. Le kilo de viande a augmenté de 3 euros de moins que le kilo de poisson. Combien coûte le kilo de viande maintenant ?

**Tc4p**

Au supermarché, le kilo de poisson a augmenté de 5 euros cette année. Un kilo de poisson coûte maintenant 12 euros. Au début de l'année, le kilo de viande coûtait le même prix que le kilo de poisson. Le kilo de viande coûte maintenant 3 euros de moins que le kilos de poisson. De combien d'euros le kilo de viande a-t-il augmenté ?

**Cc1t**

Antoine a 5 billes. Quand Antoine réunit ses billes avec celles de Paul, ils ont 12 billes ensemble. Paul réunit ses billes avec celles de Jacques. Jacques a 3 billes de moins qu'Antoine. Combien Paul et Jacques ont-ils de billes ensemble ?

**Cc1p**

Antoine a 5 billes. Quand Antoine réunit ses billes avec celles de Paul, ils ont 12 billes ensemble. Quand Paul et Jacques réunissent leurs billes, cela fait 3 billes de moins. Combien Jacques a-t-il de billes ?

**Cc2t**

Quand Médor monte sur la balance chez le vétérinaire, la balance indique 6 kilos. Quand Médor et Rex montent ensemble sur la balance chez le vétérinaire, la balance indique 15 kilos. Fido et Rex montent ensemble sur la balance chez le vétérinaire. Fido pèse 2 kilos de moins que Médor. Combien Fido et Rex pèsent-ils ensemble ?

**Cc2p**

Quand Médor monte sur la balance chez le vétérinaire, la balance indique 6 kilos. Quand Médor et Rex montent ensemble sur la balance chez le vétérinaire, la balance indique 15 kilos. Lorsque Fido et Rex montent sur la balance ensemble, la balance indique 2 kilos de moins. Combien pèse Fido ?

**Cc3t**

Dans la classe de CM2, il y a 6 élèves. Si on réunit les CM2 et les CM1, cela fait un groupe de 15 élèves. On fait un groupe réunissant les CE2 et les CM1. Dans la classe de CE2, il y a 2 élèves de moins qu'en CM2. Combien y a-t-il d'élèves dans le groupe réunissant les CE2 et les CM1 ?

**Cc3p**

Dans la classe de CM2, il y a 6 élèves. Si on réunit les CM2 et les CM1, cela fait un groupe de 15 élèves. Si on réunit les CE2 et les CM1, le groupe a 2 élèves de moins. Combien y a-t-il d'élèves en CE2 ?

**Cc4t**

Un livre coûte 9 euros. Si on achète un livre et une règle, on paie 14 euros. On achète une règle et un cahier. Le cahier coûte 2 euros de moins que le livre. Combien coûtent la règle et le cahier ensemble ?

**Cc4p**

Un livre coûte 9 euros. Si on achète un livre et une règle, on paie 14 euros. On achète une règle et un cahier. Cela coûte 2 euros de moins que lorsque l'on achète un livre et une règle. Combien coûte le cahier ?

# PROBLEMES — VERSION 2

**Tc1t**

Pendant la récréation, Lucas gagne 6 billes. Après la récréation, Lucas a 15 billes. Avant la récréation, Simon avait autant de billes que Lucas. Pendant la récréation, Simon gagne des billes, et après la récréation, il a 4 billes de moins que Lucas. Combien Simon a-t-il gagné de billes pendant la récréation ?

**Tc1p**

Pendant la récréation, Lucas gagne 6 billes. Après la récréation, Lucas a 15 billes. Avant la récréation, Simon avait autant de billes que Lucas. Pendant la récréation, Simon gagne 4 billes de moins que Lucas. Combien Simon a-t-il de billes après la récréation ?

**Tc2t**

Cette année, Théo a été pesé par le pédiatre. Théo a pris 5 kilos depuis le début de l'année. Théo pèse maintenant 14 kilos. Au début de l'année, Nicolas pesait le même poids que Théo. Maintenant, Nicolas pèse 2 kilos de moins que Théo. Combien de kilos Nicolas a-t-il pris cette année ?

**Tc2p**

Cette année, Théo a été pesé par le pédiatre. Théo a pris 5 kilos depuis le début de l'année. Théo pèse maintenant 14 kilos. Au début de l'année, Nicolas pesait le même poids que Théo. Nicolas a pris 2 kilos de moins que Théo cette année. Combien Nicolas pèse-t-il maintenant ?

**Tc3t**

En janvier, 6 enfants se sont inscrits à la chorale. Après janvier, il y a 13 enfants à la chorale. Avant janvier, il y avait autant d'enfants inscrits au football qu'à la chorale. En janvier, il y a eu de nouvelles inscriptions au football. Après janvier, il y a 2 enfants de moins au football qu'à la chorale. Combien d'enfants se sont inscrits au football en janvier ?

**Tc3p**

En janvier, 6 enfants se sont inscrits à la chorale. Après janvier, il y a 13 enfants à la chorale. Avant janvier, il y avait autant d'enfants inscrits au football qu'à la chorale. En

janvier, il y a eu 2 inscriptions de moins au football qu'à la chorale. Combien y a-t-il d'enfants au football après janvier ?

**Tc4t**

Pour Noël, Camille reçoit 7 euros. Après Noël, Camille a 12 euros dans sa tirelire. Avant Noël, Léa avait autant d'argent que Camille dans sa tirelire. Pour Noël, Léa reçoit 3 euros de moins que Camille. Combien Léa a-t-elle d'argent dans sa tirelire après Noël ?

**Tc4p**

Pour Noël, Camille reçoit 7 euros. Après Noël, Camille a 12 euros dans sa tirelire. Avant Noël, Léa avait autant d'argent que Camille dans sa tirelire. Pour Noël, Léa reçoit de l'argent. Après Noël, Léa a 3 euros de moins que Camille dans sa tirelire. Combien d'euros Léa a-t-elle reçu pour Noël ?

**Cc1t**

Antoine a 7 billes. Quand Antoine réunit ses billes avec celles de Paul, ils ont 16 billes ensemble. Jacques a 4 billes de moins qu'Antoine. Paul réunit ses billes avec celles de Jacques. Combien Paul et Jacques ont-ils de billes ensemble ?

**Cc1p**

Antoine a 7 billes. Quand Antoine réunit ses billes avec celles de Paul, ils ont 16 billes ensemble. Quand Paul et Jacques réunissent leurs billes, cela fait 4 billes de moins. Combien Jacques a-t-il de billes ?

**Cc2t**

Quand Médor monte sur la balance chez le vétérinaire, la balance indique 9 kilos. Quand Médor et Rex montent ensemble sur la balance chez le vétérinaire, la balance indique 15 kilos. Fido pèse 4 kilos de moins que Médor. Fido et Rex montent ensemble sur la balance chez le vétérinaire. Combien Fido et Rex pèsent-ils ensemble ?

**Cc2p**

Quand Médor monte sur la balance chez le vétérinaire, la balance indique 9 kilos. Quand Médor et Rex montent ensemble sur la balance chez le vétérinaire, la balance indique 15 kilos. Lorsque Fido et Rex montent sur la balance ensemble, la balance indique 4 kilos de moins. Combien pèse Fido ?

**Cc3t**

Dans la classe de CM2, il y a 9 élèves. Si on réunit les CM2 et les CM1, cela fait un groupe de 17 élèves. Dans la classe de CE2, il y a 3 élèves de moins qu'en CM2. On fait un groupe réunissant les CE2 et les CM1. Combien y a-t-il d'élèves dans ce groupe ?

**Cc3p**

Dans la classe de CM2, il y a 9 élèves. Si on réunit les CM2 et les CM1, cela fait un groupe de 17 élèves. Si on réunit les CE2 et les CM1, le groupe a 3 élèves de moins. Combien y a-t-il d'élèves en CE2 ?

**Cc4t**

Jules achète un livre à 7 euros et une règle. Jules paie 15 euros. Aurélien achète une règle et un cahier. Le cahier coûte 3 euros de moins que le livre. Combien Aurélien doit-il payer ?

**Cc4p**

Jules achète un livre à 7 euros et une règle. Jules paie 15 euros. Aurélien achète une règle et un cahier.

En tout, Aurélien paie 3 euros de moins que Jules. Combien coûte le cahier ?

# CREATION D'UN NOUVEAU MODE DE CREATION DE PROBLEME DANS DIANE

La base de données sur laquelle nous avons travaillé ne contenait pas des réponses brutes d'élèves, mais l'interprétation qu'en a fait le chercheur qui en est à l'origine. Il a, en d'autres termes, réalisé le diagnostic comportemental des réponses des élèves de manière manuelle. Un EIAH doit, dans la mesure du possible, se passer de l'intervention de l'humain dans la constitution de ces diagnostics, il est donc souhaitable qu'il dispose d'un diagnostic comportemental automatique. Pour des raisons que nous développons ci-après, il est aussi important, dans une perspective croisée de recherche en psychologie cognitive et en EIAH d'avoir un environnement qui détient une certaine flexibilité dans la création des problèmes qu'il propose. Cette partie vise à répondre à ces problèmes dans le cadre d'une amélioration de l'EIAH DIANE.

## Nouveau système de création de problèmes dans DIANE

### Concepteur d'exercice.

Face à la problématique de rendre DIANE plus pratique dans son usage dans des recherches en psychologie cognitive dans le domaine des PAEV, nous avons identifié qu'un système composé de gabarits pouvait avoir un intérêt pratique. L'enjeu, dans ces systèmes, est de pouvoir être utilisé et crée par un public non expert. Nous avons étudié la manière dont Pépite utilisait ce concept pour étendre l'ensemble des problèmes. Pour répondre à cette problématique, nous avons mis en place une interface permettant au public expert, mais non informaticien de produire des gabarits qui à leur tour permettent de générer des problèmes variés. Elle s'inspire du problème « du magicien » de PépiGen que nous avons étudié dans la partie théorique correspondante.

À première vue, l'interface que nous avons créée se présente comme un bloc de texte. Cependant, elle est enrichie par un système de balises pour permettre à DIANE de « comprendre » l'énoncé en indiquant les éléments importants (nombres de l'énoncé, question, prénoms). Ces éléments pourront, dans une autre interface, être remplacés par des valeurs différentes automatiquement. La Figure 51 présente cette interface et son utilisation. La zone d'écriture peut être enrichie par les commandes d'insertion situées au-dessous. Un élément est sélectionné et le clic sur la commande d'insertion produit le balisage tel qu'on le voit dans la zone de texte. Le cadre en pointillé situé à droite de la figure illustre la suite d'étape pour effectuer cet enrichissement dans l'interface. Pour

chaque élément ajouté, une valeur par défaut est choisie par l'utilisateur. La liste des insertions est personnalisable par le biais du bouton « ajouter ». Par le biais de ce bouton, une autre interface apparait, permettant à l'utilisateur de produire ou d'utiliser un autre type d'élément (cf. Figure 52).

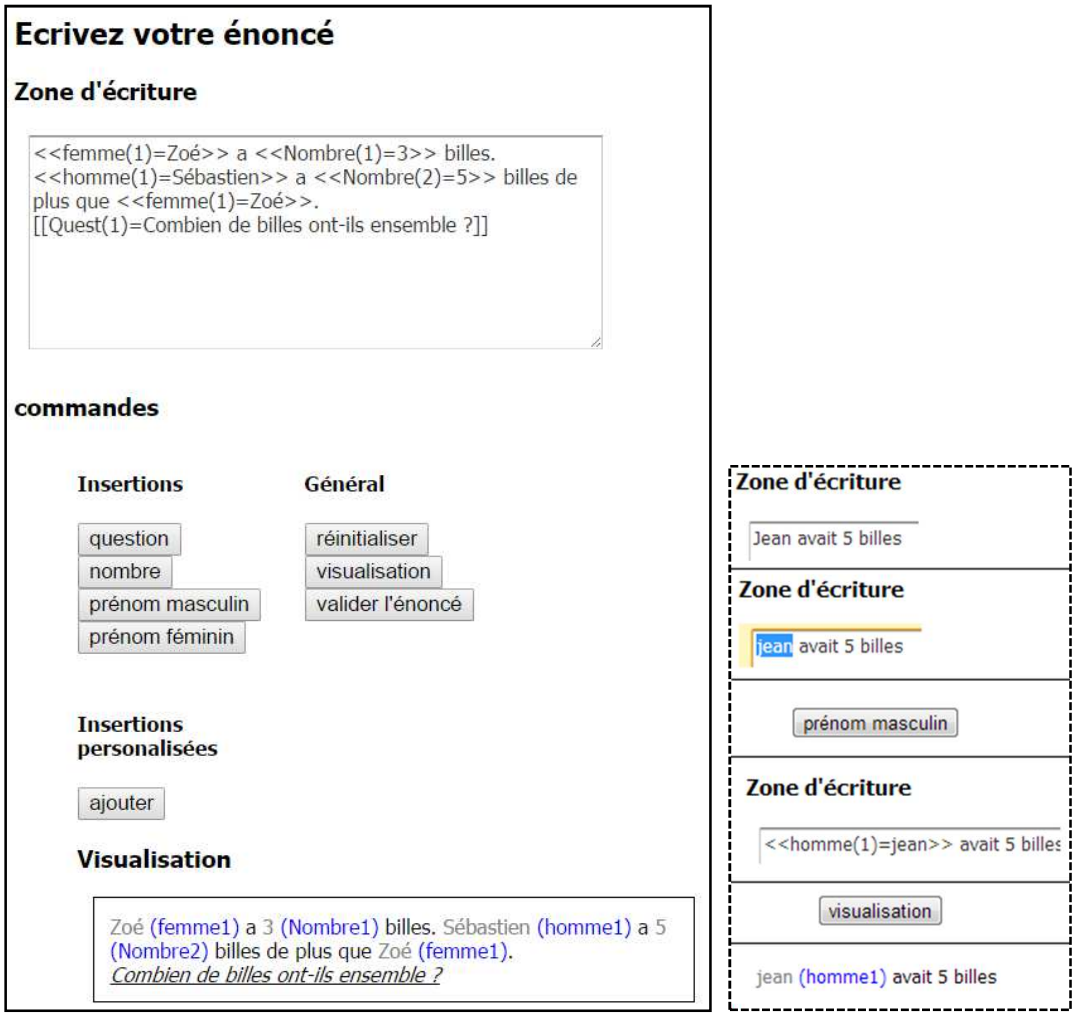



Figure 51. Conception de l'énoncé lors de la création d'un gabarit

Le balisage des éléments dans la construction d'un gabarit permet à DIANE de connaître et indexer les nombres du problème, et de savoir quels éléments peuvent être remplacés par d'autres éléments de la même catégorie. Par exemple, remplacer un prénom féminin par un autre devrait être sans conséquence sur la résolution, mais peut permettre de produire un énoncé différent.

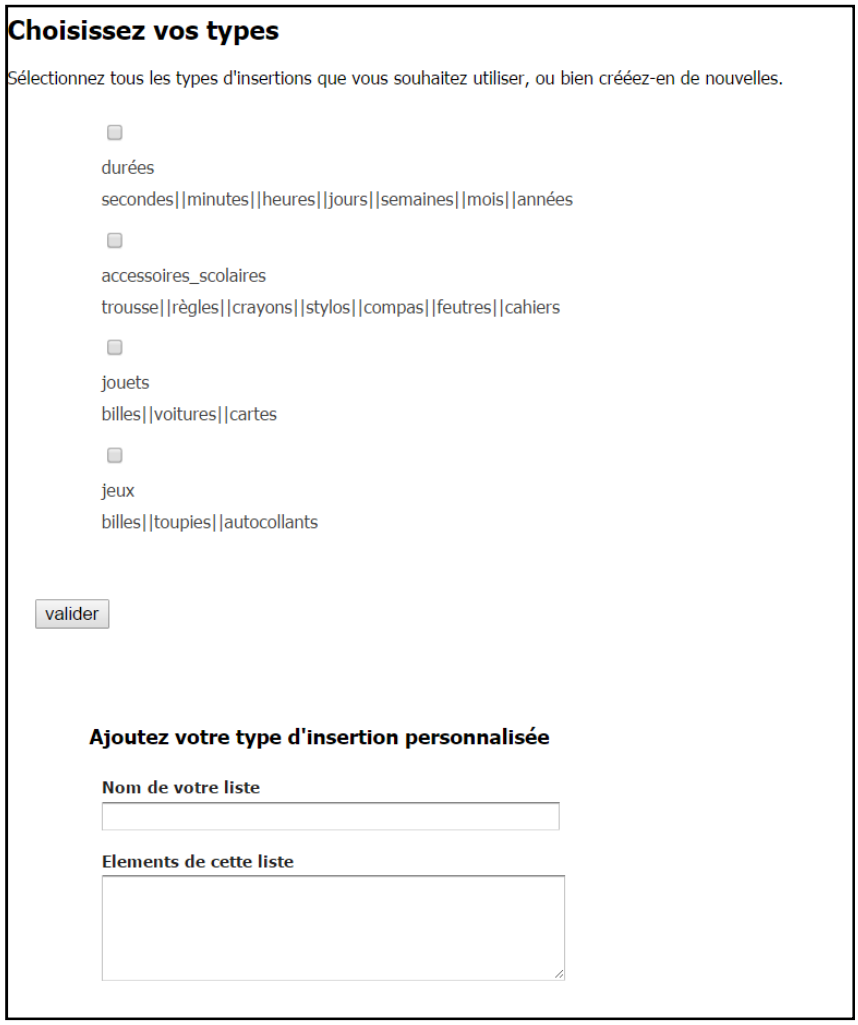



Figure 52. Interface de création et de sélection de listes personnalisées. L'utilisateur peut choisir sa liste supplémentaire ou créer la sienne.

L'interface précédente permet de créer un texte certes « enrichi », mais ne permet pas de donner aucune information portant sur l'aspect psychologique du problème. Les catégories d'appartenances et les différents facteurs influant la résolution sont absents de cette interface. Il est souhaitable de permettre au psychologue de renseigner ces informations. Nous avons insisté précédemment sur l'importance de laisser ouvert ces indicateurs, c'est-à-dire de pouvoir en ajouter de nouveaux dans le futur et les associer aux problèmes qui sont concernés. Cette fonctionnalité est rendue possible par une interface dédiée, présentée en Figure 53.

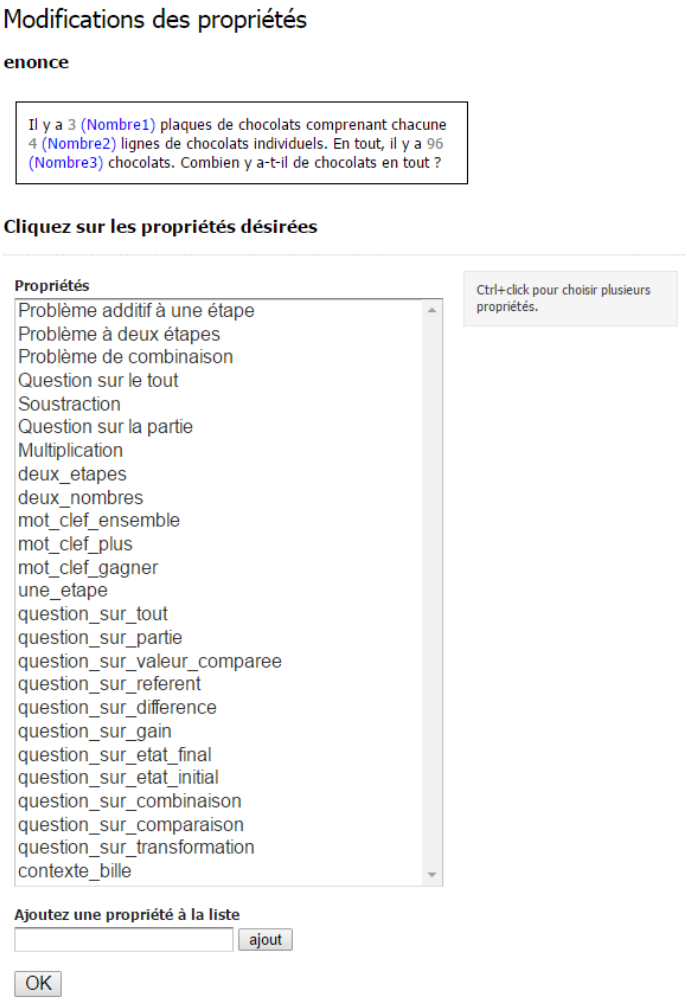



Figure 53. Ajout de propriétés à un problème dans DIANE.

### Création de la liste de réponses possibles

DIANE a été dotée d'un nouveau système de diagnostic comportemental. L'étude de ce système est l'objet principal de la partie de ce manuscrit. Or, comme nous l'avons souligné dans les parties théoriques, « généricier » le diagnostic comportemental de DIANE ne doit pas s'accompagner de la perte de la capacité du système de reconnaître certains types de réponses préétablies par le chercheur. Nous avons choisi de construire une interface permettant de spécifier ces « réponses attendues ». Ainsi, l'environnement, après avoir réalisé son diagnostic comportemental générique, peut accomplir l'étape suivante qui est d'associer la réponse de l'élève à une réponse préétablie.

Après avoir produit son problème, avoir rentré le texte de l'énoncé, l'utilisateur a donc la possibilité d'associer au problème autant de réponses possibles qu'il souhaite. Ces réponses sont décrites de la manière suivante :

- Une partie « détection » contenant la variable numérique calculée (à laisser vide si la réponse est qualitative) ainsi que des mots-clefs que pourrait contenir la réponse de l'élève.
- Une partie « propriétés » permettant d'associer des propriétés à la réponse attendue (bonne réponse, mauvaise réponse, mauvaise interprétation de la relation de comparaison, etc.). Tout comme les propriétés sur les problèmes, l'utilisateur a la possibilité d'en ajouter. Elles permettent d'avoir un jugement sur la réponse de l'élève, et d'en informer plus tard le module de diagnostic cognitif. Ce système est semblable à ce qui est réalisé dans Pépite.

Une partie du diagnostic comportemental dit « enrichi » est donc rendu auteur.

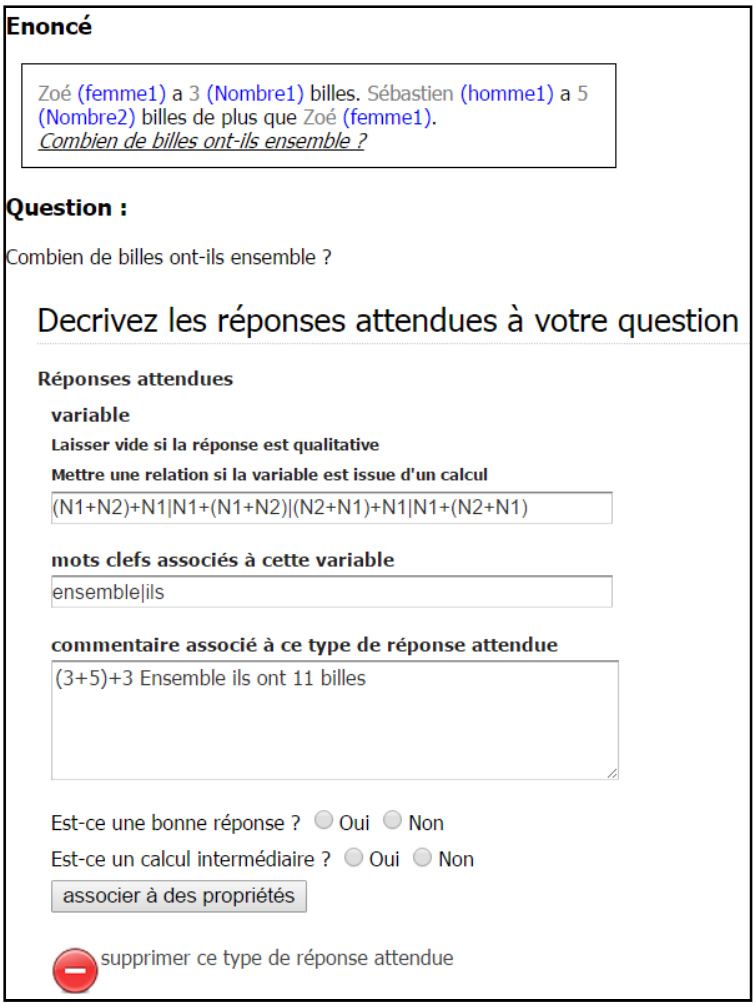



Figure 54. Création de la liste des réponses attendues.

**Fonctionnalités laissées au professeur.**

Pour des raisons de sécurité et de complexité, le professeur n'est pas autorisé à produire des gabarits dans DIANE, cependant il a accès à **l'utilisation de ces gabarits**. Il a donc la possibilité de générer des problèmes relativement variés sans perdre la puissance du diagnostic de DIANE. La Figure 55 représente l'interface laissée au professeur pour générer un problème. L'utilisateur peut cliquer sur « générer un clone de ce problème » afin d'obtenir une nouvelle proposition. Il crée le problème en cliquant sur « générer le problème"

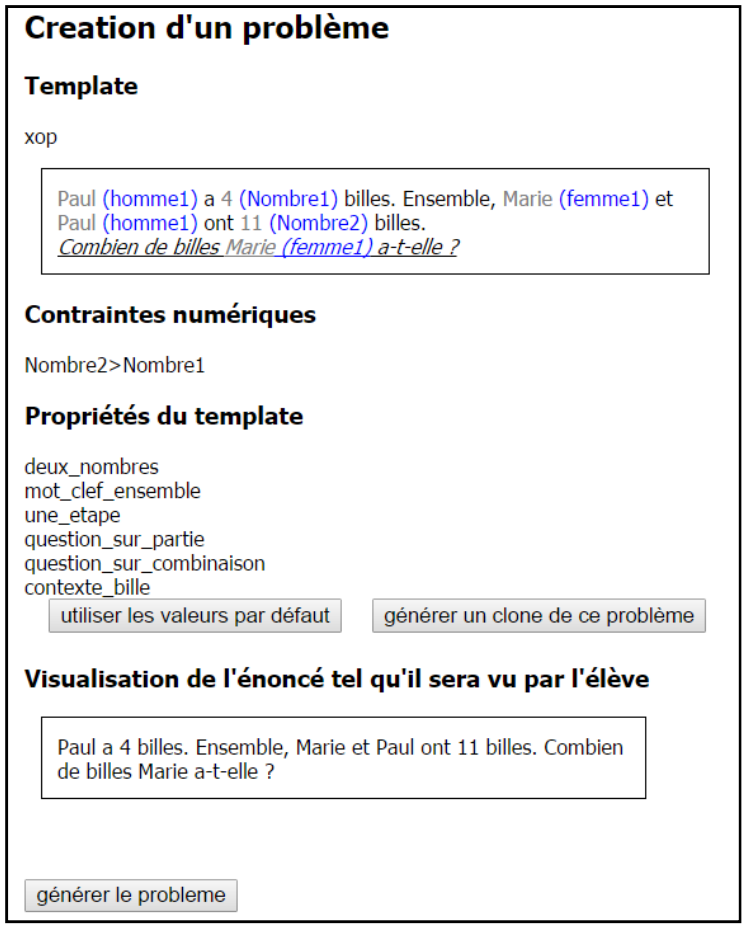



Figure 55. Utilisation d'un gabarit pour générer un nouveau problème.

Cependant, il peut également vouloir concevoir ses propres problèmes. Un mode « texte seul » a donc été créé. Dans ce cas, n'importe quel énoncé peut être utilisé dans DIANE, mais la fonctionnalité de diagnostic est limitée. Seul le diagnostic comportemental peut être produit et il est impossible d'utiliser la fonctionnalité permettant de définir des réponses attendues et donc de les détecter sur la base du diagnostic comportemental.

Notre système de conception de problème a des propriétés qui le rend unique dans la littérature : (1) les gabarits ne peuvent être créés par l'utilisateur. En effet, généralement, l'utilisateur n'a accès qu'à la partie « paramétrage » des gabarits. (2)

L'approche est **générique** : les gabarits permettent de créer n'importe quel PAEV à structure additive. (3) La conception d'un problème peut être réalisée par des modes complémentaires. Cette dernière particularité permet de compenser le temps long que peut prendre la création d'un gabarit à partir de 0 (cf. Figure 56).

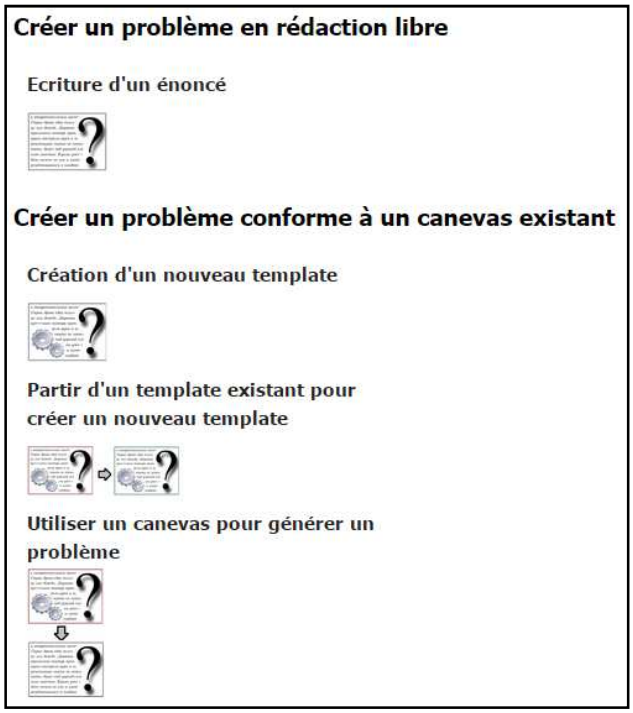



Figure 56. Différents modes de création de problèmes pour le concepteur. Dans l'interface « professeur », seuls le mode « écriture d'un énoncé » et le mode « utiliser un canevas pour générer un problème » sont disponibles.

# CONSIGNES DANS L'EXPERIMENTATION EVALUANT LE DIAGNOSTIC COMPORTEMENTAL

Bonjour,

Merci d'avoir accepté de participer à cette expérimentation.

Nous vous demandons de faire un choix forcé entre deux propositions qui représentent la formule finale donnée par l'élève. Une de ces propositions (formule 1 ou formule 2 au hasard) est issue d'un codage humain, l'autre est issue d'une procédure automatique. Nous avons rassemblé leurs désaccords et nous vous demandons de juger quel codage vous semble le bon. Par formule finale, nous entendons la formule qui détaille la provenance de ce qu'on considère être la réponse de l'élève. Nous vous demandons de mettre une croix (lettre « x ») à côté de la formule que vous pensez être la meilleure. Ce choix étant fait, nous vous demandons d'attribuer une note de confiance dans votre choix. Cette attribution se fait par une note de 1 à 4 dont nous vous donnons les différents niveaux :

1=Choix très discutable. Le cas est difficile.

2=Choix discutable, mais la formule choisie semble plus pertinente.

3=Assez sur du choix, de bonnes raisons poussent à sa sélection.

4=Absolument certain de la formule choisie. La formule choisie peut être qualifiée d'erreur.

Si vous vous trouvez en face d'une situation dans laquelle aucune formule ne vous paraît bonne, nous vous demandons quand même de prendre une décision sur celle que vous trouvez « la meilleure ». Pour vous aider dans vos décisions, les nombres de l'énoncé sont reportés dans une colonne à côté du protocole. Nous vous donnons aussi en annexe la liste des problèmes associés au code problème en première colonne.

Certaines remarques sont à prendre en compte :

- Les formules habituellement codées pour qualifier un calcul tel que 9+4+3=16 peuvent être représentées comme T1+P1+d ou (T1+P1)+d. Nous vous demandons de ne pas considérer la deuxième écriture comme une erreur.
- Lorsque la colonne finale est vide ou qu'elle contient « ininterp », on considère que la formule est déclarée ininterprétable.

- L’erreur dite « Résultat mal identifié » conserve la convention de codage à laquelle vous êtes habituée. Si 3 est le nombre calculé dans la formule 12 (P1)+3=15(T1) et que 15 est pris comme résultat, on note T1-P1 pour qualifier 15.
- Les formules représentées sous la forme T1-P1 P1-d (un espace entre deux formules) sont remplacées par la deuxième formule (P1-d) en l’occurrence.
- Les lignes sont randomisées, sauf par chance extrême, deux formules qui se suivent ne sont pas issues du même élève.

# TEST DE PERMUTATION DANS LA GENERATION D'ERREURS

modèle originel

| réponse prédite | occurrences empirique |
|---|---|
| **oui** | **15** |
| **oui** | **2** |
| non | 3 |
| non | 4 |
| non | 2 |
| non | 1 |
| non | 0 |
| non | 8 |
| non | 3 |
| non | 2 |

17 occurrences capturées

modèle permuté

| réponse prédite | occurrences empirique |
|---|---|
| **oui** | **15** |
| non | 2 |
| non | 3 |
| **oui** | **4** |
| non | 2 |
| non | 1 |
| non | 0 |
| non | 8 |
| non | 3 |
| non | 2 |

19 occurrences capturées

modèle permuté

| réponse prédite | occurrences empirique |
|---|---|
| non | 15 |
| **oui** | **2** |
| non | 3 |
| non | 4 |
| non | 2 |
| non | 1 |
| non | 0 |
| **oui** | **8** |
| non | 3 |
| non | 2 |

10 occurrences capturées

Les tests de permutation et binomiaux répondent à des logiques différentes. Le test binomial s'intéresse à la probabilité de capturer au moins 17 occurrences sur les 40 **si les réponses d'élèves étaient réparties au hasard**. Sous l'hypothèse nulle ainsi définie, la probabilité de capturer une occurrence est de 2/10. La loi binomiale nous permet donc de connaître la probabilité de capturer au moins 17 occurrences sur 40. Cette probabilité est de 0.000991, elle est largement significative.

Le test de permutation répond à la question de la probabilité suivante : **si les prédictions du modèle étaient tirées au hasard,** quelle est la probabilité de capturer au moins 17 occurrences sur les 40. Il existe 45 manières de faire 2 prédictions parmi 10. Sur ces 45 possibilités, 7 obtiennent un nombre de réponses capturées supérieur ou égal au modèle testé. La p-value est donc de 7/45=0.155, ce qui est non significatif. Plus les réponses se concentrent sur certaines catégories, plus l'écart entre les deux tests se creuse.

Quel que soit le modèle testé, les élèves ne sont pas répartis de manière uniforme sur un ensemble a priori. L'hypothèse nulle du test binomial est donc d'emblée fausse. Elle correspondrait à la théorie selon laquelle les élèves répondent au problème de manière complètement aveugle. Si, au contraire, il est reconnu que le fait que les élèves ne répondent pas au hasard, alors tester les performances d'un modèle par un test binomial est un non sens statistique car n'importe quel modèle qui tirerait ses prédictions au

hasard auraient plus de 5% de chance de passer la barre de la significativité $p < 0.05$ ce qui est paradoxal est contraire à l'esprit des statistiques inférentielles.

Au contraire, dans le test de permutation, l'hypothèse nulle repose sur la pertinence des prédictions établies et non pas sur la répartition des données. L'hypothèse nulle est définie par H0: « Le modèle tire ses k prédictions au hasard ».

# VUE PARTIELLE DES SIMULATIONS SUR LE PROBLEME TC4T

La catégorie "3C" contient trop de cas pour être contenue sur une seule page.

| Pbm | Category | Formula | Reinterpretation Model direct | Reinterpretation Model | Reinterpretation Model extended | Keyword Model | Keyword Model extended | Presence of Substraction | occurrences enfant |
|---|---|---|---|---|---|---|---|---|---|
| Tc4t | 1C | T1+d | TRUE | TRUE | TRUE | TRUE | TRUE | FALSE | 6 |
| Tc4t | 1C | T1+P1 | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 1C | P1+d | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | 1 |
| Tc4t | 1C | P1-d | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | 2 |
| Tc4t | 1C | T1-P1 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 2 |
| Tc4t | 2C | T1+(P1+d) | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 2C | (T1+d)+P1 | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 2C | (T1+P1)-d | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | 1 |
| Tc4t | 2C | (T1-d)+P1 | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE | 1 |
| Tc4t | 2C | (T1+P1)+d | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | 4 |
| Tc4t | 2C | T1-(P1+d) | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | 0 |
| Tc4t | 2C | T1-(P1-d) | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | 3 |
| Tc4t | 2C | (T1+d)-P1 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 2C | (T1-P1)-d | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | 6 |
| Tc4t | 2C | (T1-P1)+d | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | 7 |
| Tc4t | 2C | T1+(P1-d) | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | 3 |
| Tc4t | 2C | (T1-d)-P1 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | (T1-d)-d | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | 0 |
| Tc4t | 3C1R | (T1+d)-d | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | 0 |
| Tc4t | 3C1R | (T1+P1)+P1 | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C1R | (T1+d)+d | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C1R | (P1+d)+P1 | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C1R | (T1-d)+d | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | 0 |
| Tc4t | 3C1R | (P1+d)+d | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C1R | (P1+d)-d | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | 0 |
| Tc4t | 3C1R | T1+(T1-d) | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | d-(P1-d) | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | P1+(P1-d) | FALSE | TRUE | TRUE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | T1-(T1-d) | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | (T1+P1)-P1 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | (T1+P1)-T1 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | (T1+P1)+T1 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | 0 |
| Tc4t | 3C1R | T1-(T1-P1) | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | (T1-P1)-P1 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | (T1-P1)+P1 | FALSE | TRUE | TRUE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | (T1+d)-T1 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | P1-(P1-d) | FALSE | FALSE | TRUE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | (P1+d)-P1 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | d+(P1-d) | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 0 |
| Tc4t | 3C1R | T1+(T1-P1) | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | 1 |
| Tc4t | 3C1R | (T1+d)+T1 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | 0 |
| Tc4t | 3C | ((T1-d)+P1)+d | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | 0 |
| Tc4t | 3C | T1-((P1+d)-d) | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | 0 |
| Tc4t | 3C | ((T1+d)+P1)+d | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C | T1+((P1+d)+d) | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C | T1+((P1+d)-d) | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | 0 |
| Tc4t | 3C | (T1+P1)+(P1+c | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C | ((T1+P1)+d)+d | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C | ((P1+d)+P1)+T | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C | T1-((P1+d)+d) | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | 0 |
| Tc4t | 3C | (T1+(P1+d))+P | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |
| Tc4t | 3C | ((T1+d)+d)+P1 | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | 0 |

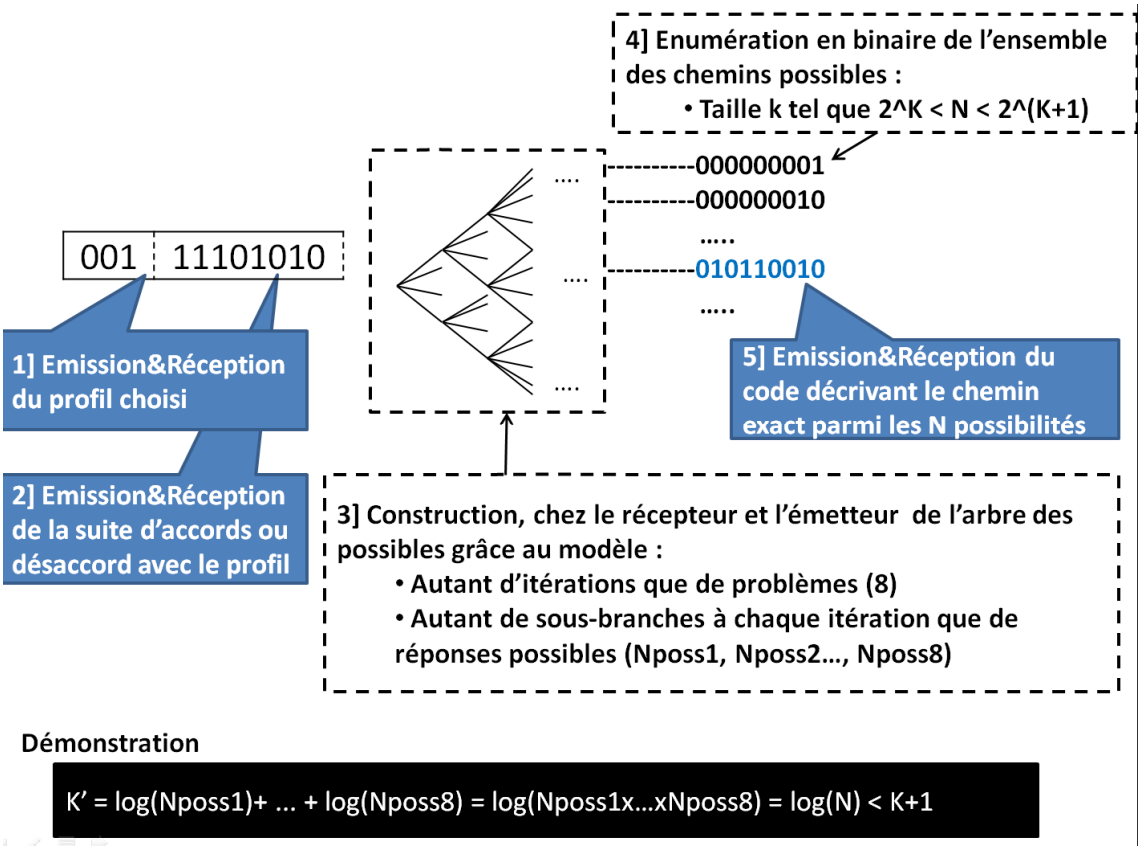
# SIMULATION DU MODELE DE REINTERPRETATIONS SUR LE PROBLEME TC4T

| Formule de calcul | Réinterprétations utilisées | Formule de calcul dont les nombres sont remplacés par les objets du problème |
|---|---|---|
| (T1-P1)+(P1-d) | [] | ((PoissonEF-PoissonGAIN)-PoissonEIminusViandeEI)+(PoissonGAIN-PoissonGAINminusViandeGAIN)=ViandeEF |
| T1-d | [] | PoissonEF-(PoissonEIminusViandeEI+PoissonGAINminusViandeGAIN)=ViandeEF |
| P1+((T1-P1)-d) | ['P1 interpreted as PoissonEI'] | (PoissonEI-PoissonEIminusViandeEI)+((PoissonEF-PoissonEI)-PoissonGAINminusViandeGAIN)=ViandeEF |
| P1-d | ['P1 interpreted as PoissonEF'] | PoissonEF-(PoissonEIminusViandeEI+PoissonGAINminusViandeGAIN)=ViandeEF |
| T1+(P1-d) | ['T1 interpreted as PoissonEI'] | (PoissonEI-PoissonEIminusViandeEI)+(PoissonGAIN-PoissonGAINminusViandeGAIN)=ViandeEF |
| (T1+P1)-d | ['T1 interpreted as PoissonEI'] | (PoissonEI+PoissonGAIN)-(PoissonEIminusViandeEI+PoissonGAINminusViandeGAIN)=ViandeEF |
| T1 | ['dEI interpreted as PoissonEFminusViandeEF'] | PoissonEF-PoissonEFminusViandeEF=ViandeEF |
| (T1-P1)+d | ['d interpreted as ViandeGAIN'] | ((PoissonEF-PoissonGAIN)-PoissonEIminusViandeEI)+ViandeGAIN=ViandeEF |
| T1-(P1-d) | ['d interpreted as ViandeGAIN'] | PoissonEF-(PoissonEIminusViandeEI+(PoissonGAIN-ViandeGAIN))=ViandeEF |

| | | |
|---|---|---|
| (T1-P1)-d | ['-d interpreted as ViandeGAIN'] | ((PoissonEF-PoissonGAIN)-PoissonEIminusViandeEI)+ViandeGAIN=ViandeEF |
| T1-(P1+d) | ['-d interpreted as ViandeGAIN'] | PoissonEF-(PoissonEIminusViandeEI+(PoissonGAIN-ViandeGAIN))=ViandeEF |
| (T1-P1)+(P1+d) | ['-d interpreted as PoissonGAINminusViandeGAIN'] | ((PoissonEF-PoissonGAIN)-PoissonEIminusViandeEI)+(PoissonGAIN-PoissonGAINminusViandeGAIN)=ViandeEF |
| T1-d | ['d interpreted as PoissonEFminusViandeEF'] | PoissonEF-PoissonEFminusViandeEF=ViandeEF |
| (T1-P1)+(P1-d) | ['d interpreted as PoissonEFminusViandeEF'] | ((PoissonEF-PoissonGAIN)-PoissonEIminusViandeEI)+(PoissonGAIN-(PoissonEFminusViandeEF-PoissonEIminusViandeEI))=ViandeEF |
| T1+d | ['-d interpreted as PoissonEFminusViandeEF'] | PoissonEF-PoissonEFminusViandeEF=ViandeEF |
| d | ['d interpreted as ViandeEF'] | ViandeEF |
| P1+(P1-d) | ['P1 interpreted as PoissonEI'] | (PoissonEI-PoissonEIminusViandeEI)+(PoissonGAIN-PoissonGAINminusViandeGAIN)=ViandeEF |
| (T1-P1)+P1 | ['dEI interpreted as PoissonGAINminusViandeGAIN'] | ((PoissonEF-PoissonGAIN)-PoissonEIminusViandeEI)+(PoissonGAIN-PoissonGAINminusViandeGAIN)=ViandeEF |
| P1+((T1-P1)-d) | ['P1 interpreted as PoissonEI'] | ((PoissonEF-PoissonGAIN)-PoissonEIminusViandeEI)+((PoissonEF-(PoissonEF-PoissonGAIN))-PoissonGAINminusViandeGAIN)=ViandeEF |

| | | |
|---|---|---|
| ((T1-P1)-d)+P1 | ['dEI interpreted as PoissonGAINminusViandeGAIN'] | ((PoissonEF-PoissonGAIN)-((PoissonEIminusViandeEI+PoissonGAINminusViandeGAIN)-PoissonGAINminusViandeGAIN))+(PoissonGAIN-PoissonGAINminusViandeGAIN)=ViandeEF |
| (T1-P1)+((T1-P1)-d) | ['P1 interpreted as PoissonEI'] | ((PoissonEF-PoissonGAIN)-PoissonEIminusViandeEI)+((PoissonEF-(PoissonEF-PoissonGAIN))-PoissonGAINminusViandeGAIN)=ViandeEF |
| ((T1-P1)-d)+(P1-d) | ['dEI interpreted as PoissonGAINminusViandeGAIN'] | ((PoissonEF-PoissonGAIN)-((PoissonEIminusViandeEI+PoissonGAINminusViandeGAIN)-PoissonGAINminusViandeGAIN))+(PoissonGAIN-PoissonGAINminusViandeGAIN)=ViandeEF |

# METHODE DE CODAGE GLOBALE POUR JUSTIFIER LA SOMME DES LOGARITHMES DANS LE CALCUL DE LA TAILLE DU MESSAGE COMPRESSE




Dans un premier temps, seul le profil et la suite d'accord/désaccord au profil sont envoyés, reçus et décodés. À partir de ces informations il est possible de construire du côté de l'émetteur et du récepteur un arbre de possibilités : « Puisqu'au problème 1, le profil est respecté, alors il y a Nposs1 possibilité de réponse qui forment autant de branches dans mon arbre »,« Puisqu'au problème 2, le profil est respecté, alors il y a Nposs2 possibilités de réponse qui forment autant de sous-branches dans mon arbre » .....

La technique consiste donc à construire un arbre qui permet d'énumérer toutes les possibilités puis d'indexer toutes les feuilles de l'arbre par un code binaire commençant par 000...001 et se terminant par le dernier élément de l'arbre. Par construction, chaque feuille détient l'information complète des réponses de l'apprenant. Par construction encore, il y a $Nposs_1$*Nposs*...$Nposs_8$ feuilles. La taille K minimale du message binaire énumérant les $Nposs_1$*Nposs*...$Nposs_8$ doit vérifier

- $2^{\wedge K} > Nposs_1 * Nposs * ... Nposs_8$ (suffisamment grand pour pouvoir contenir toutes les possibilités)

ainsi que

- $Nposs_1 * Nposs * ... Nposs_8 > 2^{\wedge}(K-1)$ (mais pas plus grand que nécessaire, si cette l'inégalité était fausse, on aurait choisi K-1 et non pas K pour énumérer $Nposs_1 * Nposs * ... Nposs_8$)

De la deuxième inégalité, il en suit que $K-1 < Log(Nposs_1 * Nposs * ... Nposs_8)$ soit que $K < Log(Nposs_1) + ... + Log(Nposs_8) + 1$.

K étant la longueur de la chaîne binaire codant l'ensemble des réponses de l'élève par cette technique globale, on a bien l'inégalité recherchée qui montre que sommer les $Log(Nposs_i)$ pour estimer la taille du message compressé ne peut réaliser une estimation à la hausse de plus d'un bit.