Metafictional anaphora

Semeijn, Merel

Published in:
Proceedings of the 2018 ESSLLI Student Session

Proceedings of the ESSLLI 2018 Student Session

30th European Summer School in Logic, Language & Information

Preface

These proceedings contain the papers presented at the Student Session of the 30th European Summer School in Logic, Language and Information (ESSLLI 2018), which was held at Sofia University "St. Kl. Ohridski" in Sofia, Bulgaria from August 6th to 17th, 2018. The Student Session is part of the ESSLLI tradition and was organized for the 30th time this year. It is an excellent venue for students to present their work on a diverse range of topics at the interface of logic, language and information, and to receive valuable feedback from renowned experts in their respective fields. The ESSLLI Student Session accepts submissions for three different tracks: Language and Computation (LaCo), Logic and Computation (LoCo), and Logic and Language (LoLa). The Student Session attracted submissions this year from all over Europe and beyond for each of the above tracks. As in previous years, the submissions were of high quality and acceptance decisions were hard to make. Of the submissions, 16 were presented as talks and 8 submissions were presented in form of a poster. Due to a special request by the author, one of the papers was not included in the online proceedings.

Four area experts, renowned in their respective fields, agreed to help in the reviewing process and support the student co-chairs of each track. We are deeply grateful for their support and help. We would also like to thank the ESSLLI Organizing Committee, especially Petya Osenova and Kiril Simov for organizing the entire summer school and supporting the Student Session in numerous ways, as well as the Program Committee chair Laura Kallmeyer. Thanks go to the chairs of the previous Student Sessions, in particular to Johannes Wahle and Karoliina Lohiniva for providing us with many of the materials from the previous years and for their advice. As in previous years, Springer has generously offered prizes for the Best Paper and Best Poster Award, and for this we are very grateful. This year we introduced an additional prize, the Axioms Award, for innovation in the fields of logic/mathematics. This award was generously provided by the Axioms Journal. Most importantly, we would like to thank all those who submitted to the Student Session, for you are the ones that make the Student Session such an exciting event to organize and attend.

Jennifer Sikos
Editor, 2018 ESSLLI Student Session Proceedings
6 August 2018

Organization Committee

Chair

Jennifer Sikos (*Universität Stuttgart*)

Language & Computation co-chairs

Martin Schmitt (*LMU Munich*)
Chantal Van Son (*Vrije U. Amsterdam*)

Logic & Language co-chairs

Carina Kauf (*Universität Göttingen*)
Swantje Tönnis (*Universität Graz*)

Logic & Computation co-chairs

Ilina Stoilkovska (*TU Wien*)
Nika Pona (*Universitat de Barcelona*)

Area Experts

Language & Computation

James Pustejovsky *(Brandeis University)*
Ivan Vulić (*University of Cambridge*)

Logic & Computation

Pavel Naumov (*Vassar College*)

Logic & Language

Jacopo Romoli (*Ulster Universität*)

Student Session Program

1st week	Monday	Tuesday	Wednesday	Thursday	Friday
	6th	7th	8th	9th	10th
	LoCo	LoLa	LaCo	LoLa	
15:50-16:20	*Social Choice and the Problem of Recommending Essential Readings* Silvan Hungerbühler, Haukur Páll Jóhnsson, Grzegorz Lisowski, Max Rapp	*Conservativeness, Language, and Deflationary Metaontology* Jonas Raab	*Playing with Information Source* Velislava Todorova	*Compositionality in privative adjectives: extending Dual Content Semantics* Joshua Martin	Beth prize talk
16:20-16:50	*Rule-based Reasoners in Epistemic Logic* Anthia Solaki	*Disjunction under Deontic Modals: Experimental Data* Ying Liu	*D3 as a 2-MCFL* Konstantinos Kogkalidis, Orestis Melkonian	*Definiteness in Shan* Mary Moroney	
Poster flash	*Explainability of irrational argument labelings* Grzegorz Lisowski	*Metafictional anaphora: A comparison of different accounts* Merel Semeijn	*Incorporating Chinese Radicals Into Neural Machine Translation: Deeper Than Character Level* Lifeng Han, Shaohui Kuang	*Free relatives, feature recycling, and reprojection in Minimalist Grammars* Richard Stockwell	

2nd week	Monday	Tuesday	Wednesday	Thursday	Friday
	13th	14th	15th	16th	17th
	LoLa	LaCo	LoLa	Laco/LoCo	
15:50-16:20	*Fighting for a share of the covers: Accounting for inaccessible readings of plural predicates* Kurt Erbach	*Classifying Estonian Web Texts* Kristiina Vaik	*Interpreting Intensifiers for Relative Adjectives: Comparing Models and Theories* Zhuoye Zhao	*The Limitations of Cross-language Word Embeddings Evaluation* Amir Bakarov	Posters
16:20-16:50	*"First things First": an Inquisitive Plausibility-Urgency Model* Zhuoye Zhao, Paul Seip	*The Challenge of Natural Language Understanding - what can Humans teach Machines about Language?* Lenka Bajčetić	*Representing Scalar Implicatures in Distributional Semantics* Maxime Corbeil	*Harrop: A new tool in the kitchen of intuitionistic logic* Andrea Condoluci, Matteo Manighetti	
Poster flash	*Towards an analysis of agent-oriented manner adverbials in German* Ekaterina Gabrovska	*Towards a Cognitive Model of the Semantics of Spatial Prepositions* Adam Richard-Bollans	*Perspective blending in graphic media* Sofia Bimpikou	*Simulating the No Alternatives Argument in a Social Setting* Lauren Edlin	Awards

Table of Contents

Language & Computation

Logic & Computation

Logic & Language

The Challenge of Natural Language Understanding - What Can Humans Teach Machines about Language?

Lenka Bajčetić

Vrije Universiteit l.b.bajcetic@student.vu.nl

Abstract. In this paper, discussing the famous Turing's test and the Chinese Room Argument, I delve into the question of what language understanding means for humans, and what it can mean for a machine. Using the "solved" problem of Word Sense Disambiguation (WSD) and IBM Watson as examples, I question the level of actual language understanding achieved with the current state-of-the-art approaches. Considering the principle with which humans successfully deal with ambiguity and understand each other, I propose a model which learns language gradually and handles open domain by asking for clarification.

Keywords: Natural Language Understanding · Symbol Manipulation · Language Ambiguity Handling

1 Introduction

Language is a complex social institution, with human communication and interaction as its primary function [Par91]. Language understanding is an internal, mental and psychological process where a person attaches a meaning to a word. It is impossible to define all the aspects this encompasses in the mind of each individual person. We cannot know how exactly another human being understands or processes something. I use the term *processes* because I want to point out that despite the immense difference in the way humans and machines process language, the same term can be used for both concepts. Actually, Natural Language Understanding is an inherently human thing and as such, quite "unnatural" to machines.

For people, language understanding requires, among other things: understanding sounds of words, talking, reading, writing, remembering, replying, but also reacting emotionally and having an internal thought process about the content of language. The memories and feelings that arise in a person from language understanding are individual and too metaphysical to be discussed in this paper. But it is hard to determine the boundaries or a definition of what is Human Language Understanding, without the human part.

Some aspects of human language understanding, are rather easily mimed. Reading and writing are default skills for computer programs, while a human child needs time and practice. Transferring text to speech and vice versa is done

with very high accuracy for English language, and soon we can expect the same for other languages as well. Many algorithms model different aspects of language syntax and semantics. Human memory can be understood as data with which an algorithm 'knows' and decides upon, and if an algorithm can logically decide upon which gaps in its knowledge to fill next, this can be seen as learning.

The question is - can machines overcome the various programming and sensory insufficiencies to leap across the difference between symbol manipulation and **understanding** their meaning?

2 How Do We Know Someone Understands?

For other human beings, we assume the ability and capacity to understand. However, if we say something to a person, let's say in a foreign country, and they ignore us, we would just assume they did not understand. Information is defined as meaningful data. So technically, we present other humans with what we think is meaningful data, and if they do not react as the meaning requires, we assume they do not understand.

While we are talking to someone, we do not question whether they understand or not, as long as they respond to what we say. The communication works because both sides have a constant awareness that there might be misunderstanding between them. The ability to distinguish these cases and solve misunderstanding is what makes people the masters of understanding. So, because people react, in a human way which we expect from them, we accept they understand language. This is not quite applicable for machines, because of the essential differences among human and computer hardware and software. However, we can expect a machine to act as close as possible to a human, in the medium which we share equally with machines - written text. This is exactly what Alan Turing has proposed.

3 Turing's Test

"Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain - that is, not only write it but know that it had written it." [Tur50]

For a machine to be considered thinking, many expect that it needs to achieve a human level of consciousness and emotions. Turing disregarded this question, even though he had admitted a paradox connected with any attempt of localizing consciousness. The mystery of consciousness and thought does not necessarily have to be solved, in order to create a machine which could pass as a human, he argues.

Because of the "polite convention that everyone thinks" [Tur50] we credit other humans with understanding capacity from the start. This is why Turing proposed a scenario, the Imitation Game, where a human interrogator is communicating with two people via text. One of them is a machine, and the interrogator

needs to determine which one. A machine which has the capacity to fool a human that it is a human as well, passes the Turing's test for consciousness.

The setting of the imitation game makes a difference between a person's physical and intellectual capacities. Making the communication through teleprompter, Turing made sure the physical aspects of language understanding, such as voice, mimics, and gestures, would not be taken into account. Accepting the machine's physical disadvantage, Turing aimed to design the test so the machine still has a fair chance. A machine that processes text in a way that humans think it understands, can be considered intelligent.

This test is still the most popular test for AI sentience, and Turing has proposed it in 1950. as a replacement for the question: Can machines think? In essence, Turing turned the problem of thought and intelligence into language understanding. However, by disregarding the differences between a human and a machine he also dodged the question of deeper, inner understanding, in the sense of attaching meaning. Turing is not concerned with the meaning within, just the output.

4 Chinese Room Argument

For many, Turing's test seems insufficient to prove a machine is thinking and understanding. Inputting text and outputting a response reasonable enough to convince a person that they are talking to a human does not seem enough to call the machine intelligent.

Supporting the sceptics, Searle gave a famous proof that Turing's test is not enough for us to accredit the program with actual language understanding [Sea80]. The proof he provided is well known as the Chinese Room Argument. He compared the program which is taking the Turing's test to himself taking the test for Chinese, being in a closed room with Chinese symbols and a rule-book. He provides the correct output for the input he gets, but he does not know the meaning of any of those symbols, as he in fact does not know Chinese. Still, he is passing the Turing's test for Chinese because the output is fooling the Chinese interrogator.

This means that Turing's test is inadequate, or at least insufficient. However, Searle is forgetting something as well. Imagining himself in this situation, Searle thinks in English, and this has nothing to do with the fact that he does not speak Chinese, or that he would pass the Chinese Turing's test using the rule-book. The Chinese Turing's test is not testing his English understanding skills. The program could, potentially, also have its language which is not Chinese, but its own. The interrogator cannot know if the person inside is thinking in English, so disproving the validity of Turing's test in this case does not mean that the machine doesn't think.

It is important to note that, if Searle was in the room, he would have thought in English because he already knew English, and he would have known it because he had been taught for several years at least. In the case of an actual program, we could dismiss the argument about Searle thinking in English in the room.

Because, simply put, at what point in time could a program have learned a language - if it was written entirely by the programmers? The fact that a program can only do what it's programmers design and implement, made some question whether Turing's test is testing the machine at all. Since the person in the Chinese room depends on the rule-book, the same way that a program depends on the programmer, a Turing test is actually testing the rule writers and their understanding of the language [Mot89]. This is true in a way. But the way that NLU models are being made is changing rapidly. Programs are becoming a product of bigger and bigger groups of programmers, even companies, using huge amounts of data. Imagine if, after a year, a new person came to the room to replace the previous one. The new person would not be as good as the one experienced with passing Chinese symbols. In the beginning, the symbols were just "squiggles" but in time patterns emerged. With experience, the person is starting to reason over the symbols, in their own way, maybe in English. But maybe they have developed some internal system for recognizing the symbols.

This is exactly what is happening with Machine Learning approaches. Thanks to Moore's law and abundance of data, we can present the person in the room with so many Chinese characters, they start to learn things. They don't learn Chinese, in the general sense, but they learn in their own way how to reply to the symbols they are given. For some tasks, like playing Go, a group of people who do not know how to play Go can make a model which plays better than any human, and learns this in 3 days starting from zero [Sil17]. Searle's argument and Motzkin's reply were written before machine learning approaches showed us the possibilities of huge data and statistics. These allow a model to go far beyond the capabilities of one person. So far, these solutions showed a lot of potential in many NLP tasks, but general language understanding is still unfeasible. Nevertheless, programs and models are now equipped with some reasoning within them not entirely made by their programmers, which can be understood as thinking "on their own".

Searle claims that the way that human brains actually produce mental phenomena cannot be reproduced solely by virtue of running a computer program [Sea80]. For me, this seems pretty clear. A machine is not a human, and a computer cannot work like a human brain. This does not mean that a computer program cannot have its own way of thinking and learning. If a program processes new information, decides upon it and learns, is it not thinking, in a computer way? It is unnecessarily anthropomorphic to expect a program to behave as a human all the time. Especially when we don't completely understand humans either. I think it is important to distinguish thinking from understanding a language. Understanding a language requires a program to use human language, but thinking can be anything, and in some situations, machines already "think" better than us.

5 State-Of-The-Art

5.1 Watson

For humans, a quiz can be considered the perfect scenario to test language understanding. Answering complicated questions, solving riddles, puzzles, and associations are all good ways to test someone's knowledge and intelligence. As such, the open domain Question Answering (QA) presents a great challenge for programmers as well.

This is why a team of IBM programmers decided to test their skills and build Watson, a machine competitor for the American TV Quiz *Jeopardy*. This quiz is particularly demanding because of high precision, accurate confidence determination, complex language, breadth of domain, and speed [FECC$^+$10]. Of course, this was no easy task. It took approximately 3 years for a team composed of 20 researchers and software engineers with a range of backgrounds in natural language processing, information retrieval, machine learning, computational linguistics, and knowledge representation and reasoning, to bring Watson's performance near human level [FECC$^+$10].

The system they have built is called DeepQA, and is described as a massively parallel probabilistic evidence-based architecture. It employs more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses. More important than any particular technique is the way that DeepQA combines these overlapping approaches so all contribute to improvements in accuracy, confidence, or speed [FECC$^+$10].

In order to compete against a human champion, the system needed to produce exact answers to complex natural language questions with high precision and speed and have a reliable confidence in its answers, in 3 seconds or less. The requirements which this implied presented a tremendous challenge for Watson's developers. Ultimately, they have succeeded, managing to tackle both the breadth of open domain and unusual word phrasing, not uncommon in *Jeopardy*.

Nonetheless, it is important to note that, even though Watson works on a wide range of topics, the questions are still rather constrained. No matter how quickly and accurately the program answers questions, it still cannot handle any unexpected input. Be that as it may, Watson is truly an amazing example of how far models can go with NLU and question answering. And also, a great example of how hard it is to grasp the notion of understanding language. Because even if a system processes questions accurately, and extracts the relevant data based on the question, and does this better than a human - we still do not attribute it the power of understanding. At a lecture at Stanford University in 2012, one of the leaders of the Watson team made the following remark as he ended his talk: "The only advantage the human contestant had over Watson was that he understood the questions" [Val07].

5.2 SenseEval

A word or a sentence is defined as ambiguous if multiple alternative linguistic interpretations can be built for it [AE07], and word sense disambiguation is the task of determining which is the correct meaning. For the task of Word Sense Disambiguation, humans need to annotate texts to represent their semantics by labeling each content word (noun, verb, adjective, and adverb) with its WordNet sense. This effort is time-consuming and energy intensive, but it seems too complicated for automation.

However, in 2004. the Senseval-3 task was to perform this tagging automatically, with the hand-tagging being used as the gold standard for evaluation. In the task, no context was provided, but it was expected that participants will make use of additional WordNet information (synset, the WordNet hierarchy, and other WordNet relations) in their disambiguation.

Anyone who has annotated at least one text, knows that this is an undeniably complicated task. And yet, all top 10 systems beat the score of the inter-annotator agreement by more than 5 points [AE07]. The human inter-annotator agreement score was in fact quite low, only 67%, probably because the annotators were not experts in the field. This shows how far these kind of tests and expectations are from the actual concept of understanding language. With tests like this, we make the humans solve computer tasks and then teach the computer how to copy as well as it can, while even humans agree on the correct output for only two-thirds of the task.

For some tasks, this approach can be good enough, because it is possible that this is just what we need - a numerical value with some percentage of certainty. This calculated value, however, has one big drawback. It is an outcome of long and complex computations we know very little about and, in case of neural networks, usually do not truly understand.

A model built like this would fail the Turing test, no matter how high the accuracy. Systems for word disambiguation based on supervised machine learning algorithms and hand-annotated data are reaching human performance, but they have still not shown a decisive difference in any application, and just as often they can hurt the performance [AE07].

The fact that stat-of-the-art models are doing better than humans in particular tasks, and we are yet miles away from general NLU, shows we have a lot to learn. Like Turing, I think that the best way to teach a machine human language, is to try to mimic the way humans understand each other. Most importantly, we should find a way to replicate the way humans deal with misunderstanding.

6 How Do Humans Deal with Ambiguity?

Human language can be characterized as a systematic relationship between form and meaning [Val07]. This relationship is rarely straightforward, because word meaning is infinitely variable and context sensitive. The fact that 121 most frequent words occupy 7.8 meanings on average shows that a lot of the time we are guessing what the other person is saying [AE07].

Unlike machines, accepting the ambiguity of language comes naturally to humans. This is because the human brain is a very powerful machine to instantly process language, and it makes sense because the human brain invented language in the first place. On the other hand, the brain invented programming languages as well, but they depend on the premise of a finite and discrete world with a limited set of rules. While solving a problem such as Word Sense Disambiguation, we assume a finite and discrete set of senses [AE07], in order to present the problem in a solvable manner. However, it is very difficult to enforce this kind of premise onto human language because of its intricate complexity. The key difference between natural and artificial languages is the fact that an artificial language can be fully circumscribed and studied in its entirety, and a natural language cannot [Gun92]. However, we can try to copy the way humans handle this complexity.

Word sense ambiguity is a trace of the fundamental process underlying language understanding. Domain constraints sense [AE07] and in an open domain we have an unlimited set of fuzzy meanings. When communicating, humans handle the open domain with ease. This is because, when processing what they have heard or read, people assume the most likely interpretation, given the choice of expression and a-priori likelihood of message [Par91]. This is known as the Principle of strategic communication, and it allows us avoid painstaking accuracy and precision in everyday communication.

In a way, the Principle of strategic communication is similar to lazy acquisition and just-in-time compiling. Lazy acquisition defers resource acquisition to the latest possible point in time during system execution, in order to optimize resource image [Kir01]. We say that a compiler works just-in-time, when it doesn't load libraries until they are actually used, to not overflow the working memory with unnecessary knowledge.

The are many benefits of lazy acquisition and just-in-time compilation: efficient resource usage with no redundancies makes the system scalable and more robust to resource exhaustion [Kir01]. Of course, these approaches have down sides too. Avoiding steps to save on time, can also lead to losing time due to bad planning and unexpected issues that can arise from omitting some knowledge we thought was not needed. Relying on handling input on the go means we need a complex system which handles unpredictability.

Why do people talk approximately? Approximate language use allows a simplified cognitive representation and a simplified inference process. For these reasons, humans accept the ambiguousness that comes along with using imprecise language. In Figure 1, we can see the reflection of this preference of impreciseness. Looking at the frequency of word usage for number words ten to twenty, we can recognize that round numbers are preferred to odd. The most commonly used are ten, twenty, twelve and fifteen. This is because people select a scale of coarseness strategically for communication. If a person does not need precise measurements, insisting on accuracy becomes counter-productive for communication. Explanations are used only when misunderstanding already exists, not before. This ability to set the coarseness appropriately to the context, but also

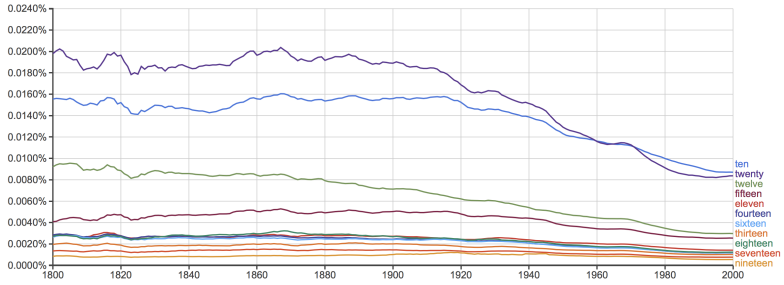


Fig. 1. Frequency of word use for number words ten to twenty, *Google N-gram*

to the level of understanding of others who are supposed to understand is the key to successful communication.

If the goal is to make a system which converses in a human-like way, I think it is important to remember that when it comes to knowledge, there are cases when less is more. Humans always balance between precision required to understand each other on the one hand, and generalization needed for efficient communication on the other. Trying to fill the gaps of our models with more and more data, is not beneficial for creating a human-like model.

7 A Different Approach

Machine learning, both supervised and unsupervised, shows promising success in solving many NLU tasks such as SensEval [AE07]. The scores are boosted by more data, more features tagged, and tuning hyper-parameters. However, the focus is on the evaluation part of the task and how the solution will perform, and not actually creating a system which understands language.

Since the goal of NLU is understanding, correctly determining the meanings of the words is fundamental. In his paper controversially titled *I don't believe in word senses* Kilgarriff focuses on WSD, saying that lexical ambiguity is perhaps the most important problem an NLU system is facing [Kil02]. If we choose to create a system which can understand semantic content, we need to solve the problem of misunderstanding arising from language ambiguity. In order to do so, we need to re-think the way we approach WSD and implement a more human-inspired algorithm.

If we look at the way humans understand each other, we see that humans are not "above" ambiguity, but they have efficient methods of resolving it. When somebody says something that we are unsure of, we check by comparing our understanding to the "truth". This is why it seems to me that creating a dialog system for handling ambiguity by asking for confirmation would be a good way for solving this issue, better than trying to find the best statistical estimate.

In this case, we need to focus on the uncertainty which triggers the question asking mechanism. This mechanism depends on the person who is talking to the machine to clarify any existing misunderstandings. This way, the machine learns language in a more organic way, solidifying previous knowledge and making sure it can still make sense, before it continues learning new things. Graduate acquisition of knowledge might be crucial to having an inner understanding of something so complex as natural language. This is similar to what Turing proposes: "Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain." [Tur50]

This idea is almost 70 years old, and yet, it has not been implemented. Perhaps because, like most concisely phrased ideas, it is in fact extremely complicated. However, it is my opinion that this approach makes the most sense as a beginning of a true General NLU system. The wonders of technology we have now, such as Neural Networks, should not be omitted from the model. But they cannot provide the core decision making, because of their lack of transparency. A model which is to pass any test for true NLU will have to be able to support its words with reasoning, which a Neural Network system cannot do.

In order to be able to provide its thought process, the program needs to know *why* it understood language the way it did. But, we do not want just a very comprehensive rule-book for handling Chinese. In order to go above this, we should allow the program to make its own rule-book, with enough time to actually *learn*. Mistakes are a normal part of learning, and we accept them as a part of our humanness, so in this process the program should be given time to make mistakes. If we teach a program how to learn and how to correct its mistakes, we can create an environment for developing thought processes through language. Allowing a computer to reason, learn, and communicate with and through natural language is what, I think, Natural Language Understanding should be.

Concluding this paper, I want to try to answer the question from the Introduction: can machines overcome their programming and sensory insufficiencies and progress from symbol manipulation to true understanding of meaning? The answer depends on our definition of what is *true* meaning. If we insist on an exact replica of a human brain in code, I would have to say no. But, an open-domain system which understands humans in a way that humans understand each other seems feasible to me.

References

[AE07] Eneko Agirre and Philip Edmonds. *Word Sense Disambiguation*. 2007.

[FECC+10] D. Ferrucci, E.Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. *Building Watson: An Overview of the Deep QA Project*. 2010.

[Gun92] Carl A. Gunter. *Semantics of Programming Languages*. MIT press, 1992.

[Kil02] Adam Kilgarriff. I don't believe in word senses. pages 2–3, 2002.

[Kir01] Michael Kircher. Lazy acquisition. pages 8–10, 2001.

[Mot89] Elhanan Motzkin. Artificial intelligence and the chinese room: An exchange. 1989.

[Par91] Prashant Parikh. Communication and strategic inference. *Linguistics and Philosophy*, 14:473–514, 1991.

[Sea80] John R. Searle. Is the mind's brain a computer program? *Scientific American*, 1980.

[Sil17] David Silver. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017.

[Tur50] Alan Turing. Computing machinery and intelligence. *Mind*, 49:433–460, 1950.

[Val07] Robert Van Valin. *From NLP to NLU*. Heinrich Heine University Düsseldorf University at Buffalo and The State University of New York, 2007.

Playing with Information Source

Velislava Todorova

Sofia University

Abstract. In this paper I present a NetLogo simulation program which models human communication with indication of information source. The framework used is evolutionary game theory. Under different initial settings the individuals in the simulation either learn to systematically indicate their information source or not. The factor of most importance seems to be the impact of one's speech behaviour on their reputation. In a community where this impact is high, the individuals who do not mark their information source lose reputation quickly and are ultimately excluded from the community. My hope is that this simulation program can help understand better the grammatical category evidentiality – the prototypical way of systematically indicating information source – and why it developed in some languages and not in others.

Keywords: Information source Simulation Evolutionary game theory Evidentiality

1 Introduction

Every language has a way of indicating the information source. If this way is a special grammatical category, it is called *evidentiality*. If it is a special use of a category with a different primary meaning, it would be rather called an *evidential strategy*. And if the marking is done by lexical means, it would be simply a *lexical expression of information source*.[1] There are even further means to indicate one's source: for example, the scientific community has developed efficient and highly conventionalized, yet not properly linguistic, ways to make bibliographical references.

I have created a simulation program[2] that models human communication with a focus on information source indication. The simulation is not meant to represent specifically the linguistic marking of information source, but its main motivation is to shed light on the possible reasons for the appearence of evidentiality in some languages and not in others.

The intuition behind the simulation scenario is that the indication of information source is connected to the reputation of speakers. In Aikhenvald's (2004, p. 359) words:

[1] For a clear distinction between the possible ways to indicate information source, see (Aikhenvald 2004, esp. Section 1.2.2).

[2] It could be viewed and downloaded from `https://github.com/SlavaTodorova/InformationSourceSimulation.git`

> In a small community everyone keeps an eye on everyone else, and the more precise one is in indicating how information was acquired, the less the danger of gossip, accusation, and so on. No wonder that most languages with highly complex evidential systems are spoken by small communities.

This article will show how reputation, and most precisely the impact of one's speech on their reputation, does indeed play a role in the development of a systematic practice of marking information source.

2 Structure of the simulation

Before the start of the simulation, the user specifies the number of individuals in the population, the number of witnesses, the level of reliability of the information and the impact of the speaker's messages on their reputation. When the simulation starts, an event takes place and some individuals witness it. The witnesses might get a wrong impression of the event,[3] but either way they search for hearers to share what they think has happened. If there are uninformed individuals near the witness and if those individuals find the reputation of the witness high enough, a conversation begins. In the conversation the speaker utters a message reporting the belief they have and, optionally, marking the information source. Hearers either believe what they have heard or not, and decide if the information should be spread further. There is again the chance of misunderstanding the message.

When the whole population has been informed (or misinformed) about the event, all individuals observe, as by providence, whether their believes and statements are true or false. On the basis of these observations their strategies (to prefer one message or another, and to rather believe or disbelieve a message) are adjusted, and their reputation levels are changed. With this a step in the simulation is completed, and a new one can start, with a new event and new witnesses.

At the end of each step of the simulation, the individuals with minimal (zero) reputation are excluded from the community and if the individuals are less than the number specified by the user, a new member is added to the community. This new member has exactly the same strategy as one (a random one) of the individuals with maximal reputation (if there are such).

[3] For the sake of simplicity, in this simulation all agents are assumed to be cooperative and benevolent. This means that there would be no liars in the community. Still, in order to bring the model closer to reality there will be a chance of misunderstanding, which will result in formulation and spread of false information.

3 The Game

3.1 Players and Moves

The simulation is a game in the sense of evolutionary game theory and Fig. 1 presents its extensive form. At the beginning Nature (Player 0)[4] gives firsthand evidence to some of the players. Firsthand evidence can be interpreted correctly or incorrectly. As it is not a conscious decision the player makes, I assume it is again Nature's choice. The player cannot be sure if the belief they formed is true or false.[5] They nevertheless have a belief and search for a hearer to share it. If a hearer is found, they would be Player 2, and Player 1, the speaker, would choose either the basic message to communicate the information, or a more complicated message, marked for information source, viz. a firsthand information message.[6] I assume that the speaker chooses a message that correctly represents their belief and the only difference in the possible messages is that one is marked for information source and the other is not. Then Player 2 decides whether to believe the information. In the end both players have some utility from the conversation: in the leaves of the tree the first number is always the speaker's utility and the second one is the hearer's.

The second branch of the game tree – the hearsay subgame – starts with Nature giving hearsay evidence to a player.[7] The player might be given a true or false piece of information, but they cannot distinguish between the two cases. They have decided according to their hearer strategy (when they were Player 2 in the firsthand evidence subgame) if they will believe or doubt the information.[8] If they believe it, it can turn out that they have misunderstood.

There are two options for the case in which Player 1 has formed a belief – they can either use the basic message or a message marked for hearsay information.[9] The firsthand information message cannot be used, as its sincerity condition requires the additional belief that the speaker has witnessed the event. There

[4] Nature is a ficticious player in the game, whose actions are those choices that do not depend on either of the two actual players.

[5] The information sets (the sets of states between which a player cannot distinguish) are represented in the tree by dotted arcs.

[6] To give an example, in English the difference between these two kinds of messages would be the distinction between "It is raining" and "I see that it is raining."

[7] The hearsay information is given to players by other players in a previous stage of the same game. However, the structure of the simulation is such that whether a player will get hearsay information, is decided together with the distribution of firsthand evidence – all the individuals who didn't receive firsthand evidence, have to eventually be informed by others.

[8] Technically, the application of the hearer strategy takes place in the previous stage of the game, when Player 1 in this second branch has been Player 2 in the first branch. However I repeat this part of the game, as it is important to distinguish between the states that result from different outcomes in the previous stage.

[9] An example from English for the difference between a message of the basic type and a message of the hearsay type would be the same as between the sentences "It is raining" and "They say it is raining."

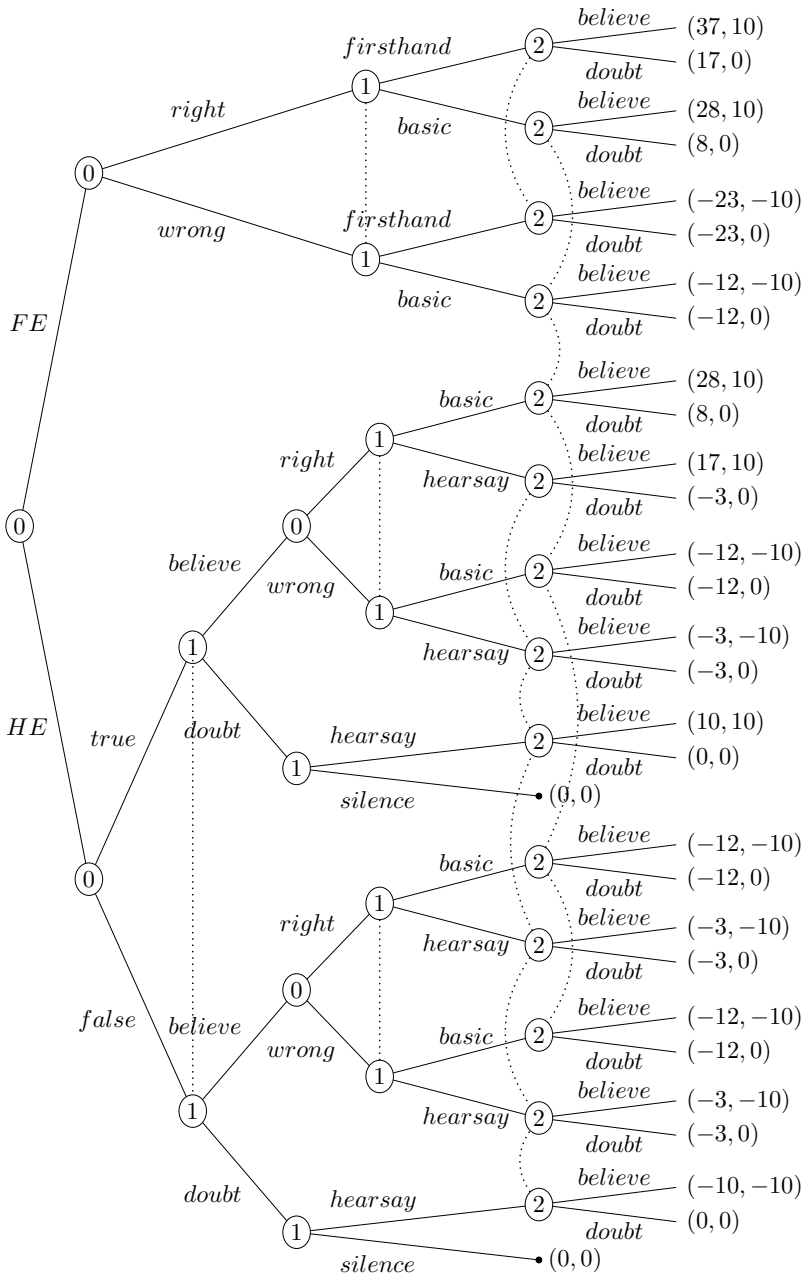


Fig. 1. Extensive form of the game. (Outcomes are calculated for reputation cost/gain values of 20 and 10 for the firsthand and the basic message respectively.)

is no possibility for unsincerity in the model (lies are not allowed), so firsthand information messages are excluded in the hearsay scenario and similarly the hearsay information messages are excluded in the firsthand evidence scenario.

In the case in which Player 1 has not formed a belief, the options are to either pass along the information using the hearsay information message, or to stay silent. The hearsay information message is the only admissible one here, as all other would not be sincere given that the speaker does not believe the information is true.[10]

Here again, just like in the firsthand evidence subgame, Player 2 has to choose if to believe the message they hear. They cannot tell if a speaker uttering the basic message was a witness or not, nor if the information this message carries is true.

There is one move by Nature that is omitted in the tree for some simplicity. After Player 2 decides to believe Player 1, it could turn out that they had formed a false belief. In this case neither the speaker nor the hearer gains or loses anything and their strategies are not updated, since neither the speaker may draw a conclusion about the persuasive power of their message, nor the hearer can blame the negative outcome of the communication on their naivity.

3.2 Outcomes, Costs and Gains

After each conversation, both the speaker and the hearer receive some utility. The precise value of the received utility depends on the perlocutionary goal the speaker had, the complexity of the message employed, the reaction of the hearer and, ultimately, on the truth of the information transmitted.

The basic outcome of the communication – the one dependent of the truth of the information – is positive for both players, if true information has been shared and believed, and negative if false information has been shared and believed. In the cases when a piece of information (true or false) is not believed, there is a neutral outcome. Table 1 presents the basic outcomes.

The basic outcome is the only factor to be considered for the hearer's utility. For the speaker there are other relevant factors. One of them is the perlocutionary goal.

In line with Martina Faller's discussion of the purposes of conversations with different evidentials in Quzco Quetchua (Faller 2006, p. 28–29), I assume that whenever the speaker does have a belief, their goal is to persuade the hearer; and that whenever the speaker shares information in the truth of which they do not believe, the goal of the communication is simply to provide the hearer with options on the basis of which they could decide for themselves what the case

[10] It is clear that in English a sentence of the form "They say it is raining" can be sincere even if the speaker is convinced it is not raining. Languages that do not use such embedding structures, but grammatical evidentiality also seem to allow for the sincere utterance of hearsay marked messages even when the speaker knows the information is false. For an example from Bulgarian, see (Smirnova 2011, p. 27) and for one from Quechua, see (Faller 2006, p. 4).

Table 1. Basic outcomes

	Player 1 (Speaker)	Player 2 (Hearer)
believed true information	10	10
not believed true information	0	0
believed false information	-10	-10
not believed false information	0	0

actually is. The gains related to the perlocutionary goals are given in Table 2. The persuading goal is only fulfilled when the hearer accepts the believe, but the alternative goal is fulfilled by the simple act of telling, and the reaction of the hearer is irrelevant.

Table 2. Perlocutionary gains for the speaker

	Perlocutionary goal:	
	Persuading	Presenting options
transferred belief	10	3
not transferred belief	0	3

Each message has an utterance cost and a conditional reputation cost, as shown in Table 3. The latter is only paid if the information turns out to be false. In case of sharing true information, there is an additional reputation gain. This aims at representing how one's utterances – according to their truth – contribute positively or negatively to one's reputation in the community.

Table 3. Costs and additional gains

	utterance cost	reputation cost	reputation gain
basic message (m_1)	2	[0, 100]	[0, 100]
firsthand message (m_2)	3	[0, 100]	[0, 100]
hearsay message (m_3)	3	[0, 100]	[0, 100]

Utterance costs are fixed, while the values of the reputation costs and gains are specified by the user (in the interval between 0 and 100). The chosen value

for the reputation costs and gains is not only used to calculate the utility of the communication, bus is also added to (or substracted form) the reputation of the speaker (which also varies between 0 and 100 and is initially 50).

The utility function for the speaker may thus be defined as follows:

$$U_s(m_i(e_j), a_k) = \begin{cases} O(m_i(e_j), a_k) - C_u(m_i) + G_r(m_i), & \text{if } e_j \text{ happened} \\ O(m_i(e_j), a_k) - C_u(m_i) - C_r(m_i), & \text{otherwise.} \end{cases} \quad (1)$$

Where O refers to the basic outcome, C_u and C_r to the utterance and reputation costs, and G_r to the reputation gain. $m_i(e_j)$ represents the uttering of a message of type m_i about event e_j. a_k for $k \in \{0, 1\}$ is the action the hearer undertakes – either to believe (a_0) or doubt (a_1) the statement.

The utility function for the hearer is considerably simpler, as it equals the basic outcome:

$$U_h(m_i(e_j), a_k) = O(m_i(e_j), a_k) \ . \quad (2)$$

The ultimate values of the utility functions of both players can be found in the game tree (Fig. 1), where the first number represents the expected utility for Player 1 (the speaker) and the second one – for Player 2 (the hearer).

4 Learning mechanism

I have chosen to model players' strategies and learning mechanisms with Pólya urns, much in the spirit of (Mühlenbernd 2011, pp. 6–8) and of the already existing Signaling Game NetLogo simulation (Wilensky 2016). Each player has a set of speaker urns for their local speaker strategies and a set of hearer urns for their local hearer strategies.[11]

There are three urns for the three speaker information sets: Ω_w for when a witness, Ω_b for when heard and believed a report and Ω_n for when the report was *not* believed. Each urn contains two kinds of balls: for each kind of message the player may choose to utter. There are other three urns for the three hearer information sets: Ω_{m_1} for the basic message, Ω_{m_2} for the firsthand information message and Ω_{m_3} for the hearsay information message. Each hearer strategy urn contains two kinds of balls: for believing the message or for discarding it. At the beginning of the game, each player's urns have the content specified in Tables 4 and 5.

After each iteration of the game, the following strategy update is made for each speaker of type t, who utters a message m, or in other words – who drew a ball b_m from the urn Ω_t at time τ to report the event e:

[11] I call *local strategy* the strategy to act in a particular way if the game has already evolved to the state in which the player has to move. Simply *strategy* will refer to a combination of local strategies and will tell us how the player would move at any point of the game.

Table 4. The initial state of the urns for the speaker strategies

	$m_1(\Omega_s)_0$	$m_2(\Omega_s)_0$	$m_3(\Omega_s)_0$	$m_{sc}(\Omega_i)_0$
$\Omega_s = \Omega_w$	100	100	0	0
$\Omega_s = \Omega_b$	100	0	100	0
$\Omega_s = \Omega_n$	0	0	100	100

Table 5. The initial state of the urns for the hearer strategies

	$a_b(\Omega_h)_0$	$a_d(\Omega_h)_0$
$\Omega_h = \Omega_{m_1}$	100	100
$\Omega_h = \Omega_{m_2}$	100	100
$\Omega_h = \Omega_{m_3}$	100	100

$$m(\Omega_t)_{\tau+1} = max[m(\Omega_t)_\tau + U_s(m(e), a), 1] \ . \tag{3}$$

Analogously, the strategy update for a hearer having drawn a ball b_a from urn Ω_m, i.e. who reacted with a to the utterance $m(e)$, would be:

$$a(\Omega_m)_{\tau+1} = max[a(\Omega_m)_\tau + U_h(m(e), a), 1] \ . \tag{4}$$

The urn cannot contain less than one ball of each type, that has been allowed in it at the beginning of the game. In this way there is always a chance for the player to change their strategy.

5 Visualization

The simulation is written in the language NetLogo,[12] and the explanation of its visualization will follow the structure of a typical NetLogo program. The basic element are the turtles,[13] these are the agents I use to represent communicating individuals. Then there are links between turtles – I represent by them the messages exchanged between the individuals.

5.1 Turtles

The turtles have shape, size, color and opacity. The shape represents the type of information source – the witnesses are square-shaped and the rest of the turtles

[12] See (Wilensky 1999).

[13] The language has been developed for simulating the behaviour of a robotized turtle, hence the extravagant name of this basic kind of agents.

have the shape of a circle. The size of the turtle represents the individual's reputation. The bigger the turtle, the greater its reputation.

Speaker local strategies are represented by color. The user can choose if they want to see the speaker local strategies for firsthand evidence, the one for believed hearsay evidence or the one for doubted hearsay evidence. In each case the probabilities of the individuals to use the three available messages (basic, firsthand information and hearsay information message) are mapped to the RGB color space. Red represents inclination towards the basic message, green – towards the firsthand information message and blue – towards the hearsay information message.

Opacity codes hearer local strategies. The user may choose the message for which to see the hearer local strategies. The turtles get the more opaque the more the individuals are inclined to believe the message. As simultaneous visualization of speaker and hearer local strategies may produce confusion, each of these visualizations can be disabled.

5.2 Links

The messages exchanged between individuals are represented with links between turtles. Color encodes type of message: red for basic message, green for firsthand information message and blue for hearsay information message. The color coding of links can be switched off.

The link is represented by a solid line if the transmitted information is true. Otherwise the line is dotted. If the hearer has believed the message, the line's opacity is the maximal possible, otherwise the opacity of the link is reduced.

6 Examples

Figures 2, 3 and 4 present the speaker strategies after 1000 communication 'steps' in a population of 100 individuals with 1 witness and reliability value of 0.9. What varies, are the values of the reputation bet for the commitment messages (the basic and the firsthand message). Each figure consists of three NetLogo views, representing the strategies for firsthand information, believed hearsay information and not believed hearsay information (in this order).

Figure 2 presents the case in which the reputation of the agents is not influenced at all by what they say and the way they say it. This is why all the dots are the same size – the agents kept their initial reputation. There seems to be a slight preference for the marked message in the firsthand information scenario and somewhat clearer preference for the unmarked message in the believed hearsay case.

Figure 3 is an example for the influence of a high reputation bet value (80 for both commitment messages). One can see how the dots are of different sizes, representing agents with different reputation levels. Furthermore, there is a clear tendency for marking hearsay information. The agents seem to have divided in their strategies towards firsthand information. In comparison with Fig. 2, here

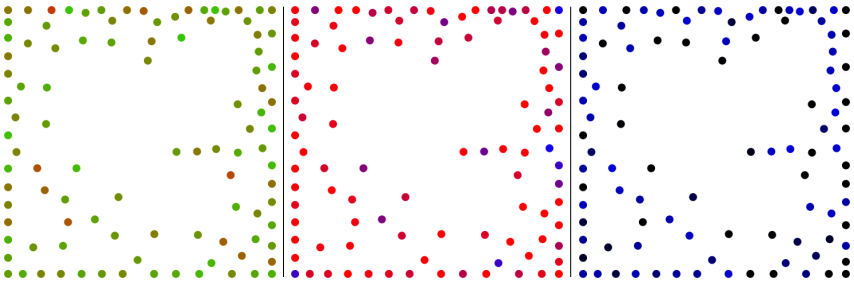


Fig. 2. Speaker local strategies for firsthand information, believed hearsay information and not believed hearsay information, with no impact of the messages on the speaker's reputation.

the speakers' preferences are clearer – they are common for the community in the case of hearsay and more a matter of personal choice in the firsthand scenario.

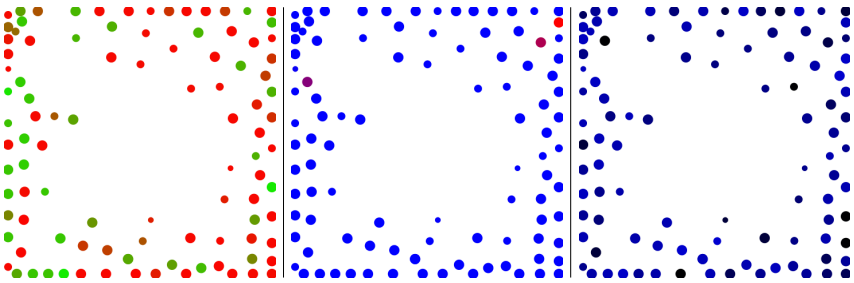
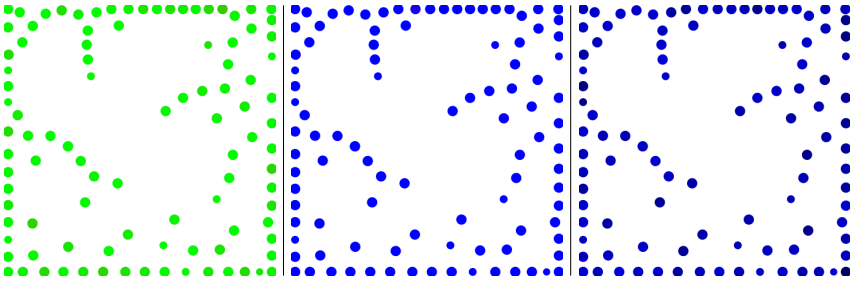


Fig. 3. Speaker local strategies for firsthand information, believed hearsay information and not believed hearsay information, with high impact of the messages (reputation bet = 80) on the speaker's reputation.

Figure 4 consists of two parts: Case A and Case B. They are two different developments that occurr when the simmulation is run twice with the same initial parameters, viz. reputaion bet value of 80 for the firsthand message and 60 for the basic.

In Case A the whole population managed to develop a strategy to mark hearsay information, as well as firsthand information. In Case B the population again developed a preference (somewhat weaker, though) for marking firsthand information, but this time they failed to adopt a strategy for marking hearsay. As a result, it is more likely for an agent in Case B to loose reputation and ultimately

Case A.



Case B.

Fig. 4. Speaker local strategies for firsthand information, believed hearsay information and not believed hearsay information, with different impacts of the commiting messages (reputation bet for firsthand message = 80, and for basic message = 60) on the speaker's reputation. Cases A. and B. are different developments of the same initial settings.

be excluded from the community, which is why there are so few agents remaining in case B, even though their initial number was 100, like in Case A.

7 Conclusion

I have described here a simulation that presents speaker reputation as one of the factors relevant for the systematic marking of information source. It was shown that the impact of one's speech on their reputation does influence the choice to indicate the information souce or not. Furthermore, we saw that in a setting with high impact of speech on reputation *not* marking hearsay information increases the risk of exclusion from the community.

The finding that reputation and systematic marking of information source are related can explain (at least to some extent) the existence of the grammatical category evidentiality in some linguistic communities. It is in line with the fact that most languages with large evidential systems are spoken in small, compact communities, where a person is very dependent on their good name.

References

Aikhenvald, Alexandra Y: Evidentiality. Oxford University Press, (2004)

Benz, Anton and Gerhard, J and van Rooij, Robert: An Introduction to Game Theory for Linguists. In: Benz, A. and Jäger, G. and Rooij, R. Van and Rooij, Robert Van: Game Theory and Pragmatics. Palgrave Macmillan UK, 1–82, (2005)

Faller, Martina: Evidentiality and Epistemic Modality at the Semantics/Pragmatics Interface. `https://www.academia.edu/25944467/Evidentiality_and_Epistemic_Modality_at_the_Semantics_Pragmatics_Interface` (2006)

Mühlenbernd, Roland: Learning with neighbours. Synthese S1, 183, 87–109 (2011)

Harsha, Prahladh: Hellinger distance, `http://www.tcs.tifr.res.in/$\sim$prahladh/teaching/2011-12/comm/lectures/l12.pdf` (2011)

Ozturk, Ozge and Papafragou, Anna: Children's Acquisition of Evidentiality. Proceedings of the 2nd Conference on Generative Approaches to Language Acquisition North America (GALANA) (2007)

Smirnova, A.: The meaning of the Bulgarian evidential and why it cannot express inferences about the future. Proceedings of SALT 21, 275–294 (2011)

Wilensky, U.: NetLogo. `http://ccl.northwestern.edu/netlogo/` Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, (1999)

Wilensky, U.: NetLogo Signaling Game model. `http://ccl.northwestern.edu/netlogo/models/SignalingGame`. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (2016)

D^3 as a 2-MCFL

Konstantinos Kogkalidis and Orestis Melkonian

University of Utrecht, The Netherlands
{k.kogkalidis,o.melkonian}@uu.nl

Abstract. We discuss the open problem of parsing the Dyck language of 3 symbols, D^3, using a 2-Multiple Context-Free Grammar. We tackle this problem by implementing a number of novel meta-grammatical techniques and present the associated software packages we developed.

Keywords: Dyck Language; Multiple context-free grammars (MCFG)

1 Introduction

Multidimensional Dyck languages[6] generalize the well-known pattern of well-bracketed pairs of parentheses to k-symbol alphabets. Our goal in this paper is to study the 3-dimensional Dyck language D^3, and the question of whether this is a 2-dimensional multiple context-free language, 2-MCFL.

For brevity's sake, this section only serves as a brief introductory guide towards relevant papers, where the interested reader will find definitions, properties and various correspondences of the problem.

1.1 Preliminaries

We use D^3 to refer to the Dyck language over the lexicographically ordered alphabet $a < b < c$, which generalizes well-bracketed parentheses over three symbols. Denoting with $\#x(w)$ the number of occurrences of symbol x within word w, any word in D^3 satisfies the following conditions:

(D1) $\#a(w) = \#b(w) = \#c(w)$
(D2) $\#a(v) \geq \#b(v) \geq \#c(v),\ \forall v \in \mathit{PrefixOf}(w)$

Eliding the second condition (D2), we get the MIX language, which represents free word order over the same alphabet. MIX has already been proven expressible by a 2-MCFG[10]; the class of multiple context-free grammars that operate on pairs of strings[2].

1.2 Motivation

Static Analysis Interestingly, the 2-symbol Dyck language is used in the *static analysis* of programming languages, where a large number of analyses are formulated as *language-reachability* problems[9].

For instance, when considering interprocedural calls as part of the source language, high precision can only be achieved by examining only control-flow paths that respect the fact that a procedure call always returns to the site of its current caller[8]. By associating the program point *before* a procedure call f_k with $(_k$, and the one *after* the call with $)_k$, the validity problem is reduced to recognizing D^2 words.

Alas, the 2-dimensional case cannot accommodate richer control-flow structures, such as exception handling via `try/catch` and Python generators via the `yield` keyword. To achieve this, one must lift the Dyck-reachability problem to a higher dimension which, given the computational cost that context-sensitive parsing induces, is currently prohibited. If D^3 is indeed a 2-MCFL, parsing it would become computationally attainable for these purposes and eventually allow scalable analysis for non-standard control-flow mechanisms by exploiting the specific structure of analysed programs, as has been recently done in the 2-dimensional case[1].

Last but not least, future research directions will open up in a multitude of analyses that are currently restrained to two dimensions, such as *program slicing*, *flow-insensitive points-to analysis* and *shape approximation*[9].

Linguistics For the characterization of natural language grammars, the extreme degree of scrambling permitted by the MIX language may be considered overly expressive[3].

On the other hand, the prefix condition of D^3 is more suggestive of free word order still respecting certain linear order constraints, as found in natural languages. Hence, it is reasonable to examine whether D^3 can also be modelled by a 2-MCFG. Such an endeavour proved quite challenging, necessitating careful study of correspondences with other mathematical constructs.

1.3 Correspondences

Young Tableaux A standard Young Tableau is defined as an assortment of n boxes into a ragged (or jagged, i.e. non-rectangular) matrix containing the integers 1 through n and arranged in such a way that the entries are strictly increasing over the rows (from left to right) and columns (from top to bottom). Reading off the entries of the boxes, one may obtain the *Yamanouchi* word by placing (in order) each character's index to the row corresponding to its lexicographical ordering.

In the case of D^3, the Tableau associated with these words is in fact *rectangular* of size $n \times 3$, and the length of the corresponding word (called a *balanced or dominant Yamanouchi word* in this context) is $3n$, where n is the number of occurrences of each unique symbol[6]. Practically, the rectangular shape ensures constraint (D1), while the ascending order of elements over rows and columns ensures constraint (D2). In that sense, a rectangular standard Young tableau of size $n \times 3$ is, as a construct, an alternative way of uniquely representing the different words of D^3. We present an example tableau in Fig.1.

a:	1	3	4	8	9	10
b:	2	5	7	11	13	15
c:	6	12	14	16	17	18

Fig. 1. Young tableau for *"abaabcbaaabcbcbccc"*

Promotions and Orbits There is an interesting transformation on Young Tableaux, namely the *Jeu-de-taquin* algorithm. When operating on a rectangular tableau $T(n, 3)$, Jeu-de-taquin consists of the following steps:

(1) Reduce all elements of T by 1 and replace the first item of the first row with an empty box $\square(x, y) := (1, 1)$.
(2) While the empty box is not at the bottom right corner of T, $\square(x, y) \neq (n, 3)$, do:
   - Pick the minimum of the elements directly to the right and below the empty box, and swap the empty box with it. $T(x, y) := min(T_{(x+1,y)}, T_{(x,y+1)})$, $\square(x', y') := (x+1, y)$ (in the case of a right-swap) or $\square(x', y') := (x, y+1)$ (in the case of a down-swap).
(3) Replace the empty box with $3n$.

The tableau obtained through Jeu-de-taquin on T is called its promotion $p(T)$. We denote by $p^k(T)$, k successive applications of Jeu-de-taquin. It has been proven that $p^{3n}(T) = T$[7]. In other words, the promotion defines an equivalence class, which we name an *orbit*, which cycles back to itself. Orbits dissect the space of D^3 into disjoint sets, i.e. every word w belongs to a particular orbit, obtained by promotions of T_w.

$\mathbf{A_2}$ *Combinatorial Spider Webs* The A_2 *irreducible combinatorial spider web* is a directed planar graph embedded in a disk that satisfies certain conditions[4]. Spider webs can be obtained through the application of a set of rules, known as the *Growth Algorithm*[7]. These operate on pairs of neighbouring nodes, collapsing them into a singular intermediate node, transforming them into a new pair or eliminating them altogether. Growth rules will be examined from a grammatical perspective in Section 2.2. Upon reaching a fixpoint, the growth process produces a well-formed Spider Web, which, in the context of D^3, can be interpreted as a visual representation of parsing a word[6,7].

A bijection also links Young Tableaux with Spider Webs. More specifically, the act of promotion is isomorphic to a combinatorial action on spider webs, namely *web rotation*[7].

Constrained Walk A Dyck word can also be visualized as a constrained *walk* within the first quadrant of $\mathbb{Z}^2$. We can assign each alphabet symbol x a vector value $\boldsymbol{v_x} \in \mathbb{Z}^2$ such that all pairs of $(\boldsymbol{v_x}, \boldsymbol{v_y})$ are linearly independent and:

$$\boldsymbol{v_a} + \boldsymbol{v_b} + \boldsymbol{v_c} = \mathbf{0} \tag{1}$$

$$\kappa\boldsymbol{v_a} + \lambda\boldsymbol{v_b} + \mu\boldsymbol{v_c} \geq \mathbf{0}, (\forall\kappa \geq \lambda \geq \mu) \tag{2}$$

We can then picture Dyck words as routes starting from $(0, 0)$. (1) means that each route must also end at $(0, 0)$ ($\cong$ (D1)), while (2) means that the x and y axes may never be crossed ($\cong$ (D2)). An example walk is depicted in Fig.2.

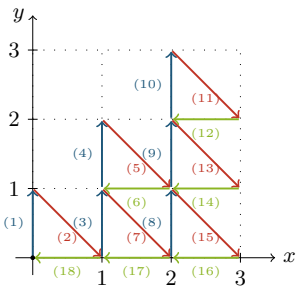


Fig. 2. The constrained walk of *"abaabcbaaabcbcbccc"* with vector value assignments $\boldsymbol{v}_a = (1, 0)$, $\boldsymbol{v}_b = (-1, 1)$, $\boldsymbol{v}_c = (0, -1)$

2 Modeling Techniques

We now present a number of novel techniques that we developed as an attempt to solve the problem at hand, incrementally moving towards more complex and abstract grammars. For the purpose of experimentation we have implemented these techniques, based on a software library for parsing MCFGs[5]. The resulting Python code is open-source and available online[1].

2.1 Triple Insertion

To set things off, we start with the grammar of *triple insertion* in Fig.3. This grammar operates on non-terminals $\mathrm{W}(x, y)$, producing $\mathrm{W}(x', y')$ with an additional triplet *a, b, c* that respects the partial orders $x < y$ and $a < b < c$. The end-word is produced through the concatenation of (x, y).

Despite being conceptually simple, this grammar consists of a large number of rules. Its expressivity is also limited; the prominent weak point is its inability to manage the effect of *straddling*, namely the generation of words whose substituents display complex interleaving patterns. Refer to Fig.10 for an example.

2.2 Meta-Grammars

To address the issue of rule size, we introduce the notion of *meta-grammars*, loosely inspired by Van Wijngaarden's work[11], which allows a more abstract view of the grammar as a whole. Specifically, we define $\mathcal{O}$ as the *meta-rule*

[1] https://github.com/omelkonian/dyck

$$\mathrm{S}(xy) \leftarrow \mathrm{W}(x, y). \tag{1}$$
$$\mathrm{W}(\epsilon, xy\mathbf{abc}) \leftarrow \mathrm{W}(x, y). \tag{2}$$
$$\ldots$$
$$\mathrm{W}(\mathbf{abc}xy, \epsilon) \leftarrow \mathrm{W}(x, y). \tag{61}$$
$$\mathrm{W}(\epsilon, \mathbf{abc}). \tag{62}$$
$$\ldots$$
$$\mathrm{W}(\mathbf{abc}, \epsilon). \tag{65}$$

Fig. 3. Grammar of triple insertions

which, given a rule format, a set of partial orders (over the tuple indices of its premises and/or newly added terminal symbols), and the MCFG dimensionality, automatically generates all the order-respecting permutations. An example of how we can abstract away from explicitly enumerating the entirety of our initial rules is showcased in Fig.4.

$$\mathrm{S}(xy) \leftarrow \mathrm{W}(x, y).$$
$$\mathcal{O}_2[\![\mathrm{W} \leftarrow \epsilon \mid \{a < b < c\}]\!].$$
$$\mathcal{O}_2[\![\mathrm{W} \leftarrow \mathrm{W} \mid \{x < y,\ a < b < c\}]\!].$$

Fig. 4. $\mathcal{G}_0$: Meta-grammar of triple insertions

This approach enhances the potential expressivity of our grammars as well. For instance, we can now extend the previous grammar with a single meta-rule that allows two non-terminals $\mathrm{W}(x, y)$, $\mathrm{W}(z, w)$ to interleave with one another, producing rearranged tuple concatenations and allowing some degree of straddling to be generated:

$$\mathcal{G}_1 : \mathcal{G}_0 + \mathcal{O}_2[\![\mathrm{W} \leftarrow \mathrm{W}, \mathrm{W} \mid \{x < y,\ z < w\}]\!].$$

The addition of this rule gets us closer to completeness, but we are still not quite there. We have thus far only used a single non-terminal, not utilizing the expressivity that an MCFG allows. To that end, we propose non-terminals to represent incomplete word *states*; that is, words that either have an extra symbol or miss one. The former are *positive* states, whereas the latter are *negative*. The inclusion of these extra states would allow for more intricate interactions.

Interestingly, there is a direct correspondence between these non-terminals and the nodes of Petersen's growth algorithm[7]. Fig.5 depicts the growth rules in the exact same web form as proposed by Petersen, modulo node branding. A

subset of these web-reduction rules are, in fact, precisely modelled by the meta-grammar $\mathcal{G}_2$ presented in Fig.6. In section 4, we briefly explain our inability to model the whole set of rules with a 2-MCFG, hence rendering our grammar complete.

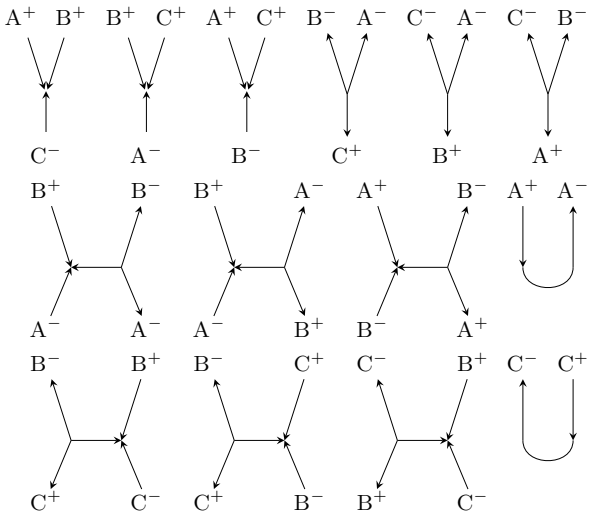


Fig. 5. Growth rules

$\mathcal{G}_2$ consists of base cases for positive states, possible state interactions, closures of pairs of inverse polarity and a universally quantified meta-rule that allows the combination of any incomplete state with a well-formed one (i.e. non-terminal W).

A further extension can be achieved through universally quantifying the notion of triple insertion, which is unique in the sense that it can insert three different terminals, each at a different position:

$$\mathcal{G}_3 : \mathcal{G}_2 + \forall\ \mathrm{K} \in \{\mathrm{A}^{+/-},\ \mathrm{B}^{+/-},\ \mathrm{C}^{+/-}\} : \mathcal{O}_2[\![\mathrm{K} \leftarrow \mathrm{K} \mid \{x < y,\ a < b < c\}]\!].$$

2.3 Rule Inference

The improved performance of the above approaches again proved insufficient to completely parse D^3. Our meta-rules are over-constrained by imposing a total order on the tuple elements, due to their inability to keep track of where the extra character(s) is. To overcome this, we split each state into multiple position-aware, *refined* states. Doing so revealed a vast amount of new interactions, as

$\mathrm{S}(xy) \leftarrow \mathrm{W}(x, y).$
$\mathcal{O}_2[\![\mathrm{W} \leftarrow \epsilon \mid \{a < b < c\}]\!].$
$\mathcal{O}_2[\![\mathrm{A}^+ \leftarrow \epsilon \mid \{a\}]\!].$
$\mathcal{O}_2[\![\mathrm{B}^+ \leftarrow \epsilon \mid \{b\}]\!].$
$\mathcal{O}_2[\![\mathrm{C}^+ \leftarrow \epsilon \mid \{c\}]\!].$
$\mathcal{O}_2[\![C^- \leftarrow A^+, B^+ \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![B^- \leftarrow A^+, C^+ \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![A^- \leftarrow B^+, C^+ \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![A^+ \leftarrow C^-, B^- \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![B^+ \leftarrow C^-, A^- \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![C^+ \leftarrow B^-, A^- \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![W \leftarrow A^+, A^- \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![W \leftarrow C^-, C^+ \mid \{x < y < z < w\}]\!].$
$\forall\ \mathrm{K} \in \{\mathrm{A}^{+/-},\ \mathrm{B}^{+/-},\ \mathrm{C}^{+/-}\}:$
$\quad \mathcal{O}_2[\![\mathrm{K} \leftarrow \mathrm{K}, \mathrm{W} \mid \{x < y,\ z < w\}]\!].$

Fig. 6. $\mathcal{G}_2$: Meta-grammar of incomplete states

evidenced by the below alteration to the original A^+, B^+ interaction (where y can now occur after z or w):

$$\mathcal{O}_2[\![C^- \leftarrow A^+_{left}, B^+ \mid \{x < y, x < z < w\}]\!].$$

In order to accommodate the interactions between this increased number of states, we need to keep track of both internal and external order constraints. At this point, the abstraction offered by our meta-grammar approach does not cover our needs any more. The same difficulty that we had encountered before is prominent once more, except now at an even higher level.

As a solution to the aforementioned limitation, we propose a system that can automatically create a full-blown m-MCFG given only the states it consists of. To accomplish this, we assign each state a unique *descriptor* that specifies the content of its tuple's elements. Aligning these descriptors with the tuple, we can then infer the descriptor of the resulting tuple of every possible state interaction. For the subset of those interactions whose resulting descriptor is matched with a state, we can now automatically infer the rule.

Formally, the system is initialized with a map $\mathcal{D}$, such as the one illustrated in Fig.7. Its domain, $dom(\mathcal{D})$, is a set of *state identifiers* and its codomain, $codom(\mathcal{D})$, is the set of their corresponding *state descriptors*.

$$
\begin{aligned}
\mathrm{W} &\mapsto (\epsilon, \epsilon) \\
\mathrm{A}_l^+ &\mapsto (a, \epsilon) \\
\mathrm{A}_r^+ &\mapsto (\epsilon, a) \\
&\vdots \\
\mathrm{C}_r^- &\mapsto (\epsilon, ab) \\
\mathrm{C}_{l,r}^- &\mapsto (a, b)
\end{aligned}
$$

Fig. 7. Map $\mathcal{D}$ for refined states

Algorithm 1 ARIS: Automatic Rule Inference System

```
procedure ARIS(𝒟)
    for X ↦ (d_1, ..., d_n) ∈ 𝒟 do
        yield X(d_1, ..., d_n).
    for X, Y ∈ dom(𝒟)^2 do
        (X_ord, Y_ord) ← (x < y < ..., z < w < ...)
        for (d_1, ..., d_n) ∈ 𝒪_2⟦_ ← X, Y | {X_ord, Y_ord}⟧ do
            for S' ∈ ELIMINATE((d_1, ..., d_n), 𝒟) do
                yield S'(d_1, ..., d_n) ← X, Y.

procedure ELIMINATE((d_1, ..., d_n), 𝒟)
    for matches ∈ ALL_ABC_TRIPLETS(d_1, ..., d_n) do
        for i ∈ 0 ... n/3 do
            for S' ∈ REMOVE_ABC_TRIPLETS(matches, i) do
                if S' ∈ codom(𝒟) then
                    yield S'
```

Meta-grammars accelerated the process of creating grammars, by letting us simply describe rules instead of explicitly defining them. ARIS builds upon this notion to raise the level of abstraction even further; one needs only specify a grammar's states and its descriptors, thus eliminating the need to define rules or even meta-rules.

3 Tools & Results

3.1 Grammar Utilities

We have implemented the modelling techniques described in Section 2 and distributed a Python package, called **dyck**, which provides the programmer with a *domain-specific language* close to this paper's mathematical notation. To facilitate experimentation, our package includes features such as grammar selection,

time measurements, word generation and soundness/completeness checking. The following example demonstrates the definition of $\mathcal{G}_1$:

```
from dyck import *
G_1 = Grammar([
    ('S <- W', {(x, y)}),
    O('W', {(a, b, c)}),
    O('W <- W', {(x, y), (a, b, c)}),
    O('W <- W, W', {(x, y), (z, w)})
])
```

3.2 Visualization

As counter-examples began to grow in size and number, we realised the necessity of a visualization tool to assist us in identifying properties they may exhibit. To that end, we distribute another Python package, called **dyckviz**, which allows the simultaneous visualization of tableau-promotion and web-rotation (grouped in their corresponding equivalence classes). An example of a web as rendered by our tool is given in Fig.8.

Young tableaux in an orbit are colour-grouped by their column indices, which sheds some light on how the *jeu-de-taquin* actually influences the structure of the corresponding Dyck words. Interesting patterns have began to emerge, which still remain to be properly investigated.

3.3 Grammar Comparisons

Fig.9 displays three charts, depicting the number of rules, percentage of counter-examples and computation times of each of our grammars for D_n^3 with n ranging from 2 to 6 (where n denotes the number of *abc* triplets). Even though none of our proposed grammars is complete, we observe that as grammars get more abstract, the number of failing parses steadily declines. This however comes at the cost of rule size growth, which in turn is associated with an increase in computation times. What this practically means is that we are unable to continue testing more elaborate grammars or scale our results to higher orders of n (note that $\|D_n^3\|$ also has a very rapid rate of expansion[2]).

4 Road to Completeness

To our knowledge, no other attempt has come so close to modelling D^3 with a 2-MCFG. We attribute this to the combination of a pragmatic approach with results from existing theoretical work. In this section, we present a collection of additional ideas, which we consider worthy of further exploration.

[2] https://oeis.org/A000108

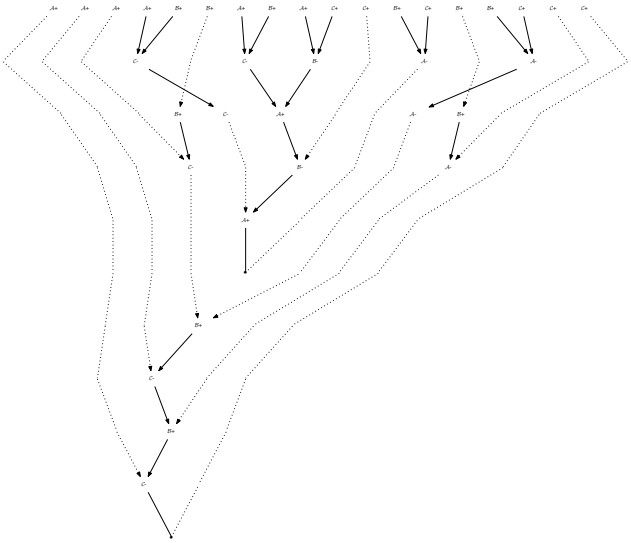


Fig. 8. Spider web of *"abaacbbacbabaccbcc"*

First-Match Policy and Relinking Possibly the most intuitive way of checking whether a word w is part of D_n^3 is checking whether a pair of links occur that match a_i to b_i and b_i to c_i $\forall i \in n$. We call this process of matching the *first-match policy*. The question arises whether a grammar can accomplish inserting a triplet of a, b, c, that would abide by the first-match policy. If that were the case, it would be relatively easy to generalize this ability by induction to every $n \in \mathbb{N}$. Unfortunately, the answer is seemingly negative; the expressiveness provided by a 2-MCFG does not allow for the arbitrary insertions required. On a related note, being able to produce a word state $W(x, y)$ where $w = xy$ and x any possible prefix of w, gives no guarantee of being able to produce the same word with an extra triplet inserted due to the straddling property.

However, if rules existed that would allow for match-making and breaking, i.e. match *relinking*, an inserted symbol could be temporarily matched with what might be its first match-policy in a local scope, and then relink it to its correct match when merging two words together.

Growth Rules Although $\mathcal{G}_2$ comes close to realizing the growth algorithm, not all of the growth rules can be translated into a 2-MCFG setting. It would be an interesting endeavour to attempt to model the element-swapping behaviour of these rules that produce two output states, without resorting to more expressive formalisms (e.g. context-sensitive grammars).

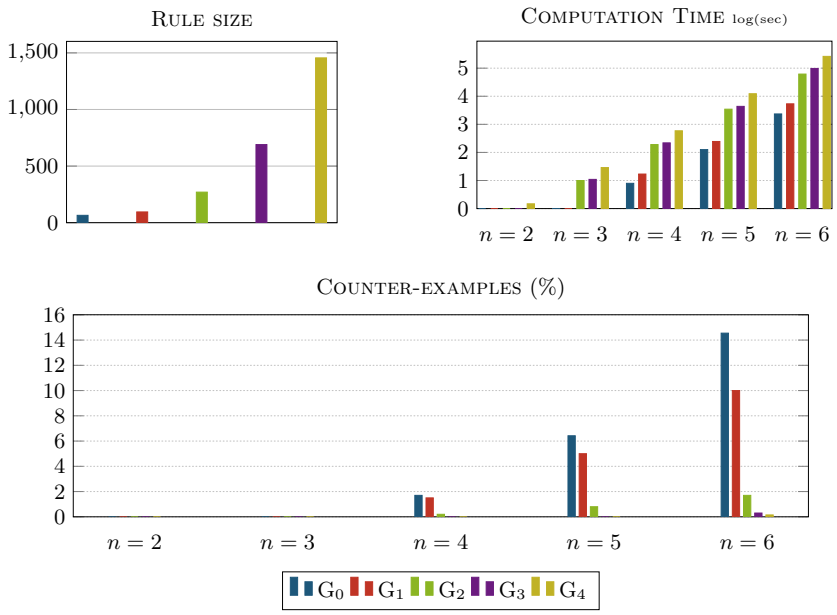


Fig. 9. Performance measures

Fig. 10. First-match policy for *"ababacbcabcc"*

Insights from promotion An interesting question is whether promotion can be handled by a 2-MCFG (as a *context-free rewriting system*). If so, it could be worth looking into the properties of orbits, to test for instance if there are promotions within an orbit that can be easier to solve than others. Solving a single promotion and transducing the solution to all equivalent words could then be a guideline towards completeness.

5 Conclusion

We tried to accurately present the intricacies of D^3 and the difficulties that arise when attempting to model it as a 2-MCFL. We have developed and introduced some novel techniques and tools, which we believe can be of use even outside the problem's narrow domain. We have largely expanded on the existing tools to accommodate MIX-style languages and systems of meta-grammars in general.

Despite our best efforts, the question of whether D^3 can actually be encapsulated within a 2-MCFG still remains unanswered. Regardless, this problem has been very rewarding to pursue, and we hope to have intrigued the interested reader enough to further research the subject, use our code, or strive for a solution on her own.

Acknowledgements

We would like to thank Dr. Michael Moortgat for introducing us to the problem, providing insightful feedback and motivating us throughout the process, as well as Dr. Jurriaan Hage for suggesting the use of multi-dimensional Dyck languages in static analysis.

References

1. Chatterjee, K., Choudhary, B., Pavlogiannis, A.: Optimal dyck reachability for data-dependence and alias analysis. Proceedings of the ACM on Programming Languages **2**(POPL), 30 (2017)
2. Götzmann, D.N.: Multiple context-free grammars
3. Kanazawa, M., Salvati, S.: Mix is not a tree-adjoining language. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. pp. 666–674. Association for Computational Linguistics (2012)
4. Kuperberg, G.: Spiders for rank 2 lie algebras. Communications in mathematical physics **180**(1), 109–151 (1996)
5. Ljunglöf, P.: Practical parsing of parallel multiple context-free grammars. In: Workshop on Tree Adjoining Grammars and Related Formalisms. p. 144 (2012)
6. Moortgat, M.: A note on multidimensional dyck languages. In: Categories and Types in Logic, Language, and Physics. pp. 279–296 (2014)
7. Petersen, T.K., Pylyavskyy, P., Rhoades, B.: Promotion and cyclic sieving via webs. Journal of Algebraic Combinatorics **30**(1), 19–41 (2009)
8. Reps, T., Horwitz, S., Sagiv, M.: Precise interprocedural dataflow analysis via graph reachability. pp. 49–61. ACM (1995)
9. Reps, T.W.: Program analysis via graph reachability. In: Logic Programming, Proceedings of the 1997 International Symposium, Port Jefferson, Long Island, NY, USA, October 13-16, 1997. pp. 5–19 (1997)
10. Salvati, S.: Mix is a 2-mcfl and the word problem in z^2 is solved by a third-order collapsible pushdown automaton. Journal of Computer and System Sciences **81**(7), 1252–1277 (2015)
11. van Wijngaarden, A.: The generative power of two-level grammars. In: ICALP (1974)

Classifying Estonian Web Texts

Kristiina Vaik

Institute of Estonian and General Linguistics, University of Tartu
kristiina.vaik@ut.ee

Abstract. The Web has become an important language resource for NLP. However, automatically crawled corpora have some shortcomings: lots of data but the content is unknown. Most research has focused on classifying multiple genres, but this paper describes the binomial classification of Estonian Web texts. The goal was to classify texts into two categories, i.e. texts following standard written language norms and texts following non-standard written language norms. Classification models were built using different supervised machine learning algorithms and BoW as features. 10-fold cross-validation was used to measure the quality of these classification models, best result was achieved by the multilayer perceptron achieving over 0.99 on accuracy. The results of classifying the manually labelled test set show that neural networks yet again outperformed other machine learning algorithms, achieving over 0.7 on accuracy.

Keywords: automatic classification, Web corpus, standard vs non-standard written language, Estonian

1 Introduction

Increasingly, corpus linguists and language technologists are turning to the Web as a source of language data. It is freely available and offers a big variety of texts. Besides traditional standard written texts (e.g. news, fiction) there is also noisy user-generated content containing lots of variation and not meeting the standardized language norms. The multitude of traditional and new genres on Web has initiated a lot of tasks in NLP. One such task is *automatic Web genre identification* or *classification*.

Automatic Web genre classification is a complex task. Different researchers have developed a variety of genre classification schemes which are mainly done using two approches: *bottom-up* and *top-down* [5,17]. In bottom-up approach researchers ask a group of people to come up with a genre classification, e.g. [4,2,6]. In top-down approach researchers use existing classifications or rely on their own knowledge of genres in order to define genre classes, e.g [27,3,15,28]. Some combine both bottom-up and top-down approaches [21]. There is no universal classification, because there is a lack of consensus about how to define *genre* and *genre classification*. Genres are recognized, yet there is a lot of disagreement in definitions, boundaries and level of specificity.

In automatic genre classification feature selection has also a very important role to play. Features used for genre classification can be grouped into three main categories – *structural* (i.e. PoS tags and *n*-grams, e.g. [13,18,28,1,22]), *lexical* (i.e. common words [26,20,22], function words [1], word *n*-grams [9,7,23,18,27], character *n*-grams [11,23]) and *other features* (i.e. keywords from the URL [29,16]). In addition to structural and lexical features text statistics (e.g. type and token ratio, average word and sentence length, frequency of punctuation marks, frequency of HTML tags, e.g. [12,14,18,22]) can also be beneficial in discriminating genres.

While most research on text classification has focused on classifying multiple genres, the objective of this paper is to describe the preliminary experiments on doing a simplified version of the standard text classification – a binomial classification of the Estonian Web corpus called the etTenTen13[1] [10]. The corpus has been semi-automatically[2] classified into seven genres (*government*, *periodical*, *informative*, *religion*, *forum*, *blog*, *unknown*). Although these texts have already been classified, the quality of this classification is questionable – third of these texts are labelled as *unknown*. Hence, there is a new situation: there is a large corpus, but the content is unknown.

The long-term aim is to create a new classification or modify the existing one, however prior to that the Estonian Web texts need to be classified into two categories: either following or not following *standard* written language norms. The reason behind it is very practical – there are no classification tools for Estonian that can discriminate noisy textual data from well-written texts. These tools are needed for checking whether there is a need for text normalization and also allows to choose appropriate preprocessing methods (paragraphs, sentences, words, morphological analysis, PoS). For example, the etTenTen has been annotated by parsers that were trained on standard written texts, therefore the quality of the annotation (e.g. morphological analysis) is a bit dubious.

There has been some research on classifying texts based on being formal or informal [24,25] and blog or non-blog [19], but that research is not directly relevant to this work. Therefore, to the author's knowledge, this is the first approach on classifying texts as following or not following standard written language norms. The aim of this paper is to find out whether it is even possible to do this kind of discrimination and what kind of classification models are best suited for this task. This classification is done by using supervised machine learning approach. These supervised models were trained on two corpora: the Balanced Corpus of Estonian [3] (representing the standard written language) and a subset from

[1] EtTenTen13 is a collection of Estonian texts automatically crawled from the Web in 2013.

[2] , Classification was based on: 1) domain classification done by the Institute of the Estonian Language (EKI), 2) information in the URLs name (e.g. an URL containing a word *comments* was classified as a *forum* type t,ext), 3) text representation (e.g. frequent appearance of times and dates or word *Vasta* (Reply) was classified as a *forum* type text), 4) domains containing at least 400 000 words were classified manually.

[3] http://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=en

the Estonian Reference Corpus called the New Media[4] (representing the non-standard written language). Ideally, the test and training set should follow the same probability distribution. However, since there isn't a corpus which has been classified into two categories (texts following standard written language norms and texts following non-standard written language norms), these corpora were used because they have been manually compiled, meaning their language usage is well-known. As the aim is to classify Web texts, a test set was composed by randomly selecting texts from the etTenTen13 corpus.

This paper presents the results of the preliminary experiments on the binary classification and discusses how to improve the quality of the classification models.

2 Datasets and Methods

Datasets. For this study three different corpora were used: the Balanced Corpus of Estonian, a subset from the Estonian Reference Corpus called the New Media and the etTenTen13. The core difference between these corpora is that the genres in the Balanced Corpus and New Media are known, while the classification in the etTenTen remains mistrustful. Balanced Corpus consists of fiction, journalistic and science texts, New Media consists of 4 subsets: chatroom, newsgroups, forums and comments. For this study chatroom texts were left out because of its distinctive language. Classification models were trained using Balanced Corpus as an example of a standard written language and New Media as an example of a non-standard written language. A subset of the etTenTen was used to compose a test set in order to validate the quality of the classification models.

Preprocessing. For noise reduction the data had to be preprocessed and normalized. `Xml` tags and line breaks in all documents were removed, certain strings were replaced with a placeholder string: web addresses with `<hyperlink>`, email addresses with `<email>` and numerals with `<number>`. Finally each document was labelled as standard or non-standard. As a result each document contained label, filename and text data columns. Commonly all punctuation marks are removed and all characters are lowercased, but for this task it seemed counterproductive. It is partly because on the Internet (e.g. comments, chatrooms) people tend to ignore the capitalization rules and it is common to use repeated punctuation marks for exaggeration (e.g. *??!!!*, *!!!*), or conversely not using punctuation at all. Thus using capitalization and interpunctuation could be helpful features distinguishing if a text follows standard or non-standard written language norms.

Features and classifiers. For setting a baseline word unigrams, i.e. Bag-of-Words (BoW) were chosen as features. BoW is essentially a very simplistic method, yet it produces surprisingly good results, e.g. [23,15]. Classification models were built using the *scikit-learn* machine learning library. Different supervised machine learning algorithms were applied: 1) multinomial Naive Bayes, 2)

[4] http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/index.php?lang=en

logistic regression, 3) linear support vector, 4) random forest and 5) multi-layer perceptron.

Procedure. 10-fold cross-validation was used to measure the qualities of these classification models. In every cycle the learning algorithm will go through these steps: 1) texts are divided into training and test set; 2) training and test set texts are converted into word vectors by varying the minimal term frequency[5] ($\min_{df}$), which are converted into frequency vectors using TF-IDF (*term frequency–inverse document frequency*) normalization; 3) training set is used for creating a classification model for given learning algorithm; 4) the model labels every text in the test set with two possible categories: standard or non-standard; 5) the accuracy measure is saved.

3 Results

3.1 Description of the Training Data, Classification Results of the First Experiment

Table 1: Training corpus

	Subset	Number of docs per subset	Total doc count	Total token count (in millions)	Avg doc length (in tokens)
New Media	forums	197	338	13.6	40314
	comments	77			
	newsgroups	64			
Balanced Corpus	fiction	138	414	12.9	31215
	science	138			
	journalism	138			
TOTAL			752	26.5	35765

The description of the training data is shown in Table 1. For training a total of 752 texts were used. The standard written language was represented by 414 fictional, journalistic and scientific texts from the Balanced Corpus (in total 55%). The non-standard written language was represented by 338 forum, comment and newsgroup type texts from the New Media subset (in total 45%).

Table 2 presents the accuracy results of the classification models obtained from the 10-fold cross-validation. It can be seen from the data that the prediction accuracy for all models is suprisingly good: regardless of the minimal term frequency, there is at least a 90% probability that a text will be labelled with the correct category by all models. The top performer was the multi-layer perceptron with an accuracy up to 99.7%. The aim of this experiment was to

[5] It is used to remove terms that appear infrequently, e.g. if $\min_{df} = 2$, ignore tokens appearing in less than 2 documents.

Table 2: Accuracy results of the cross-validation

Model	df=1	df=2	df=3	df=4	df=5	df=6	df=7	df=8
MNNB	0.940	0.949	0.941	0.943	0.942	0.943	**0.953**	0.952
LR	0.983	**0.988**	0.986	0.984	0.987	0.985	0.983	0.985
LSVC	0.992	0.992	0.991	0.992	0.992	0.992	**0.995**	0.991
RF	0.956	0.962	**0.976**	0.974	0.969	0.965	**0.976**	0.974
MLP	0.996	**0.997**	**0.997**	0.996	0.996	0.996	0.996	0.996

Notes: MNNB = multinomial Naive Bayes, LR = logistic regression, LSVC = linear support vector, RF = random forest, MLP = multi-layer perceptron.

see whether this combination of features and training set will be suitable for Estonian. Although the BoW method is simplistic, based on these results it still produces good results and could be applied for classifying Estonian texts.

3.2 Description of the Manually Labelled Test Set and Classification Results of the Second Experiment

This second experiment sets out to use these classification models which were built during the first experiment in order to find out how well these models perform on classifying the target input, i.e. the Estonian Web texts.

Manually Labelled Test Set. For evaluation a subset of Estonian Web texts from the etTenTen13 corpus was used as a test set. It consisted of 220 texts which were chosen randomly, all categories from the original genre classification were present (see Table 3).

Table 3: Overview of the test set

Genre categories of the existing corpora	Number of documents	Percentage in the test set
periodicals	74	34
informative	38	17
unknown	14	6
government	17	8
religion	17	8
forum	26	12
blog	34	15
Total	220	100

In NLP reliably annotated dataset always plays an important role. The results of research based on unreliable annotation can be considered as untrustworthy and doubtful. In order to measure the reliability of annotation, different

annotators judge the same data and the observed agreement and the Fleiss' kappa for measuring the inter-coder agreement are calculated for their judgements.

The test set was manually annotated into two categories – standard and non-standard. Manual annotation was done by three annotators. Every text was labelled by each annotator, i.e. each text got three labels. The annotator had to choose one category from a set of labels (*standard*, *non-standard*, *could not determine*). The final category for each text depended on the majority voting scheme, e.g if a text is labelled as *standard*, *standard* and *could not determine*, then the final category will be *standard*. If there was a disagreement between annotators (e.g text got judged as *standard*, *non-standard* and *could not determine*), then a fourth person was included for the final decision.

The annotators were given some instructions, but the exact amount of non-standard features (deviating ortography etc) a text had to contain in order to be labelled as non-standard was not given. The decision relied on the raters' intuition. For instance a text should be labelled as non-standard if a) it contains orthographic errors, b) the author of the text is ignoring the spelling conventions (e.g not capitalizing the first word of a sentence or the first letter of a proper name, ignoring the punctuation rules), c) the author of the text abbriviates words (e.g *krt* & *kurat* "damn it", *pmst* & *põhimõtteliselt* "in principle", *nv* & *nädalavahetus* "weekend"), d) it contains particles that are common for spoken language (e.g *aaa*, *hmm*, *mkm*) and e) it contains loanwords that are common for spoken language (e.g *poindile pihta saama* "to get the point", *khuul* "cool").

The observed and inter-annotator agreement measures of the manual annotation are depicted in Table 4. Overall, a substantial reliability with the observed agreement of 72% and Fleiss' kappa [8] of 0.656 was achieved. Both measures were also computed for each original genre category in order to identify the most and least agreed-on genre class. Fleiss' kappa values for the blog, unknown, informative and periodical categories illustrate substantial or moderate agreement among the raters, i.e. accordingly 0.531, 0.586, 0.589 and 0.689.

The results for the rest of the original genres was a bit alarming. The Fleiss' kappa values for religion and forum categories illustrate a slight agreement, i.e. 0.077 and 0.118. The kappa value for the government type texts is below zero, i.e. -0.043, indicating that there is no agreement between the raters. But if looking closely at Table 4, one can see that the observed agreement for the government category is the highest, i.e. 88%. Then, why is the inter-annotator agreement so low? These results imply that if one of the categories is much more likely than the other, then some probablities are going to be higher and this in turn will make the Fleiss' kappa much lower. In other words, government type texts are more likely to follow standard written language norms[6] which means that the data is highly homogeneous in that sense and the lack of heterogeneity produces a lower

[6] In the end 100% of the texts originating from the government category were labelled as following standard written language norms

kappa score. The same applies for the forum[7] and religion[8] genre categories. The aim of this paper is not about redoing the semi-automatic classification, the aim of this part is just to describe the test set according to its original environment.

Table 4: The observed and inter-annotator agreement of the manual annotation

Genre in etTenTen13	Observed agreement	Inter-annotator agreement
blog	65	0.531
forum	66	0.118
government	88	-0.043
informative	69	0.589
periodical	80	0.689
religion	69	0.077
unknown	64	0.586
OVERALL	72	0.656

According to the annotators' feedback there were some issues. For instance, it was difficult to classify texts that were originally labelled as forum because these texts are heterogeneous consisting of posts and replies from different users (with each user having their own writing preferences). In addition, there were texts in which the standard language content was followed by the non-standard content (e.g. news text which was followed by the comments section). This has nothing to do with the text *per se*, but rather because the XML parsing was not done properly. In fact, it is very difficult to decide firmly if a text belongs to a certain category or not. How many orthographic and spelling errors does it take for a text to be categorized as non-standard? It would be wiser to handle texts as a continuum where some texts are more and others less standard. However, for this study it was necessary to firmly divide these texts into two different categories.

Classification Results of the Second Experiment. This second experiment uses those classification models which were built during the first experiment in order to find out how well these models perform on classifying the target input. All parameters were the same as in the first experiment. The results of classifying the manually labelled test set are shown in Table 5.

In summary, compared to the results of the first experiment there is a clear tendency of decreasing accuracy results in the second experiment (see Table 2 and 5). From the data in Table 5, it can be seen that the random forest was the

[7] In the end 95% of texts originating from the forum category were labelled as following the non-standard written language norms

[8] In the end 100% of the texts originating from the religion category were labelled as following the standard written language norms

Table 5: Accuracy results on the Manullay Labelled Test Set

Model	df=1	df=2	df=3	df=4	df=5	df=6	df=7	df=8
MNNB	0.677	0.686	0.686	0.691	0.695	0.695	0.705	**0.709**
LR	**0.645**	0.595	0.600	0.595	0.595	0.605	0.595	0.595
LSVC	0.591	**0.65**	0.645	0.645	0.641	0.645	0.645	0.641
RF	0.486	0.509	0.523	0.509	0.514	0.527	**0.595**	0.500
MLP	0.727	0.727	0.727	0.727	0.732	**0.736**	**0.736**	**0.736**

Notes: MNNB = multinomial Naive Bayes, LR = logistic regression, LSVC = linear support vector, RF = random forest, MLP = multi-layer perceptron.

most poorly performing learning algorithm with the accuracy between 0.486–0.595, i.e. every other text was labelled into the wrong category. Support vector machine and logistic regression performed equally poorly – every third text was labelled into the wrong category. Yet again, the best performer was the multi-layer perceptron with an accuracy of 0.736 (min_{df}=6), i.e. it labelled 3 out of 4 texts into the right category.

Since the accuracy metric alone can be misleading if the number of observations per category in the training set are unbalanced (see Tabel 1), a confusion matrix and other evaluation metrics (i.e. precision, recall and F-score) for the best performing learning algorithm, i.e. MLP model, will be presented. Figure 1 shows how the evaluation metrics differ throughout the min_{df} value. As you can see, the results are very stable, but for each metric there is a slight increase with the rise of the min_{df}.

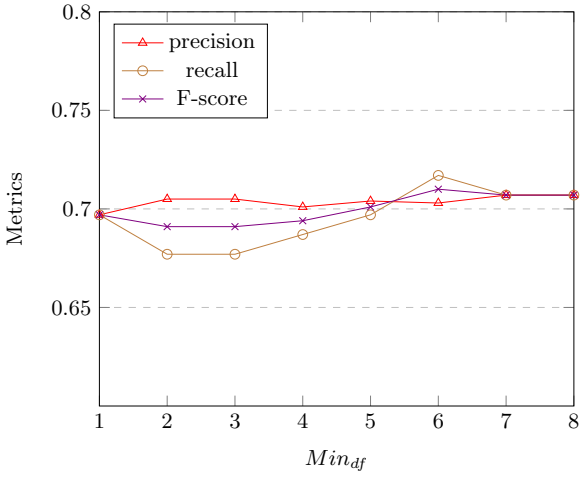


Fig. 1: Evaluation metrics for MLP

Tabel 6 presents the prediction results while the min_{df} equals to 6 while the accuracy was 0.736 (see also Table 5). The idea behind presenting the confusion matrix is to see in which cases the MLP model does a wrong decision. If we look at the Table 6, the total of texts originally categorized as standard is 120 (91+29) and the total of texts originally categorized as non-standard is 100 (29+71). The total of texts that should have been categorized as *standard* but weren't is 29, which means that 25% (29 texts) of texts were incorrectly classified into the *non-standard* category (false positives) and 75% (91 texts) were classified correctly (true postives). The total of texts that should have been categorized as *non-standard* but weren't is 29, which means that 29% (29 texts) of texts were incorrectly classified into the *standard* category (false negatives) and 71% (71 texts) were classified correctly (true negatives). This means that the MLP model isn't biased towards one or the other category since the false negatives and false positives have roughly the same percentage.

Table 6: Confusion matrix for MLP model

	Predicted: standard	Predicted: non-standard
Actual: standard	TP = 91	FN = 29
Actual: non-standard	FP = 29	TN = 71

Overall, based on the second experiment it can be assumed that the language diversity in the etTenTen13 can not be modelled using that kind of a training set, i.e. the Balanced Corpus of Estonian and the New Media corpus. Also, there is a possibility that the decreased performance is because the user-generated content of the Estonian Web corpus is a bit different from the content of the New Media corpus (texts originating from the years of 2000–2008). It has been witnessed that language of the Estonian Web is shifting into being more like the standard written language. These results further support this idea.

4 Conclusions and Future Work

Corpus linguists and language technologists are turning to the Web as a source of language data which is freely available and offers a big variety of texts. Next to traditional standard written texts (e.g. news, fiction) there is also noisy user-generated content which does not meet the standardized language norms and contains a lot of variation. This has aroused an interest in automatic Web genre classification. While most research on automatic classification has focused on classifying multiple genres, but the objective of this paper is to do a binary classification: classify texts as following or not following standard written Estonian language norms. The reason behind it is practical – there are no classification

tools for Estonian that can discriminate noisy textual data from well-written texts. These tools are needed for checking whether there is a need for text normalization and also allows to choose appropriate preprocessing methods.

The aim of the article is to describe and present the results of the automatic classification of Estonian Web texts which is a type of supervised learning problem that aims to categorize texts into a set of predefined categories based on the labelled training data. This paper evaluates the quality of the different classification models on the training set and manually labelled test set.

These models were trained on the Balanced Corpus of Estonian (example of standard written language) and New Media corpus (example of non-standard written language). To test the classification models it was necessary to compose a manually labelled subcorpus of Estonian Web texts. Classification models were built by using different supervised machine learning algorithms and BoW as features. The results obtained from the preliminary experiments show that neural networks outperformed other supervised machine learning algorithms, achieving over 0.7 on accuracy.

These results are good, but in order to increase the performance of the classifiers adding new structural, lexical features and text statistics (e.g. POS count, sentences per paragraph, words per sentence, uppercase and lowercase letters per sentence etc) should be tested. The best model, neural network classifier, achieved an accuracy of 0.99 on training set, but on test set it only achieved little over 0.73. This suggests that future work requiers a bigger and more appropriate training set. Ideally, the test and training set should follow the same probability distribution. However, since there isn't a corpus which has been classified into two categories (texts following standard written language norms and texts following non-standard written language norms), these corpora were used as a training set because they have been manually compiled – hence, their language usage is known.

The manual labelling task showed that the transition from being standard written Estonian to non-standard is very smooth. At the moment models produce a score between 0 and 1, indicating belonging to a class or not. Therefore, the classification model should be tuned to be predictive, i.e. instead of giving a discrete value, its output should be the probability score. Then, the text will be categorized based on that score.

The long-term aim is to create a new classification or modify the existing one (i.e. the etTenTen classification), however prior to that the Estonian Web texts need to be classified into two categories: either following or not following standard written language norms. The rest of the classification will be built on that.

References

1. Argamon, S., Koppel, M., Avneri, G.: Routing documents according to style. In: In Proceedings of First International Workshop on Innovative Information Systems (1998)

2. Asheghi, N., Sharoff, S., Markert, K.: Crowdsourcing for web genre annotation. Language Resources and Evaluation **50(3)**, 603–641 (2016)
3. Berninger, V., Kim, Y., Ross, S.: Building a document genre corpus: A profile of the krys i corpus. In Proceedings of the BCS-IRSG workshop on corpus profiling (2008)
4. Crowston, K., Kwaśnik, B., Rubleske, J.: Problems in the use-centered development of a taxonomy of web genres. Genres on the Web: Computational Models and Empirical Studies pp. 69–84 (2011)
5. Crowston, K., Kwasnik, H.B.: A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. Proceedings of the 37th Hawaii International Conference on System Sciences (2004)
6. Egbert, J., Biber, D., Davies, M.: Developing a bottom-up, user-based method of web register classification. Journal of the Association for Information Science and Technology **66(9)**, 1817–1831 (2015)
7. Finn, A., Kushmerick, N.: Learning to classify documents according to genre. Journal of the American Society for Information Science and Technology **57(11)**, 1506–1518 (2006)
8. Fleiss, L.J.: Measuring nominal scale agreement among many raters **76**, 378–382 (1971)
9. Freund, L., Clarke, C.L.A., Toms, E.G.: Towards genre classification for ir in the workplace. In: Proceedings of the 1st International Conference on Information Interaction in Context. pp. 30–36. IIiX, ACM, New York, NY, USA (2006). https://doi.org/10.1145/1164820.1164829, http://doi.acm.org/10.1145/1164820.1164829
10. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The tenten corpus family. In: 7th International Corpus Linguistics Conference CL 2013. pp. 125–127. Lancaster, UK (2013)
11. Kanaris, I., Stamatatos, E.: Webpage genre identification using variable-length character n-grams. In: 19th IEEE International Conference on Tools with Artificial Intelligence. vol. 2, pp. 3–10 (2007)
12. Karlgren, J., Bretan, I., Dewe, J., Hallberg, A., Wolkert, N.: Iterative information retrieval using fast clustering and usage-specific genres. In: Paper presented at the Eighth DELOS Workshop: User Interface in Digital Libraries. pp. 85–92 (1998)
13. Karlgren, J., Cutting, D.: Recognizing text genres with simple metrics using discriminant analysis. In: Proceedings of the 15th Conference on Computational Linguistics - Volume 2. pp. 1071–1075. COLING '94, Association for Computational Linguistics, Stroudsburg, PA, USA (1994). https://doi.org/10.3115/991250.991324, https://doi.org/10.3115/991250.991324
14. Kessler, B., Numberg, G., Schütze, H.: Automatic detection of text genre. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics. pp. 32–38. EACL '97, Association for Computational Linguistics, Stroudsburg, PA, USA (1997). https://doi.org/10.3115/979617.979622, https://doi.org/10.3115/979617.979622
15. Laippala, V., Luotolahti, J., Kyröläinen, A.J., Salakoski, T., Ginter, F.: Creating register sub-corpora for the finnish internet parsebank. In: Proceedings of the 21st Nordic Conference on Computational Linguistics. pp. 152–161. Association for Computational Linguistics, Gothenburg, Sweden (2017)
16. Lim, C.S., Lee, K.J., Kim, G.C.: Automatic genre detection of web documents. In: Natural Language Processing – IJCNLP 2004. pp. 310–319. Springer Berlin Heidelberg (2005)

17. Mehler, A., Sharoff, S., Santini, M.: Genres on the Web: Computational Models and Empirical Studies. Text, Speech and Language Technology, Springer Netherlands (2010)
18. Meyer Zu Eissen, S., Stein, B.: Genre Classification of Web Pages, pp. 256–269. Springer Berlin Heidelberg (2004)
19. Pardo, F., Padilla, A.: Detecting blogs independently from the language and content. pp. 1–5 (2009)
20. Petrenz, P., Webber, B.: Robust cross-lingual genre classification through comparable corpora (2012)
21. Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., Liu, X.: Genre based navigation on the web **4**, 4013 (January 2001)
22. Santini, M.: Automatic identification of genre in the web pages. Phd thesis, University of Brighton (2007)
23. Sharoff, S., Wu, Z., Markert, K.: The Web Library of Babel: evaluating genre collections. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) LREC. European Language Resources Association (2010)
24. Sheikha, A.F., Inkpen, D.: Automatic classification of documents by formality. pp. 1–5 (2010)
25. Sheikha, A.F., Inkpen, D.: Learning to classify documents according to formal and informal style. Linguistic Issues in Language Technology **8(1)**, 1–29 (2012)
26. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Text genre detection using common word frequencies. In: Proceedings of the 18th Conference on Computational Linguistics - Volume 2. pp. 808–814. COLING '00, Association for Computational Linguistics, Stroudsburg, PA, USA (2000). https://doi.org/10.3115/992730.992763, https://doi.org/10.3115/992730.992763
27. Stubbe, A., Ringsletter, C.: Recognizing genres. An Abstract Proceedings of the Colloquium Towards a reference corpus of web genres (2007)
28. Vidulin, V., Luštrek, M., Gams, M.: Using genres to improve search engines. Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing pp. 4–45 (2007)
29. Vidulin, V., Luštrek, M., Gams, M.: Multi-label approaches to web genre identification. Journal for Language Technology and Computational Linguistics **24(1)**, 97–114 (2009)

Incorporating Chinese Radicals Into Neural Machine Translation: Deeper Than Character Level

Lifeng Han[1*] and Shaohui Kuang[2*]

[1] ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
lifeng.han3@mail.dcu.ie
[2] NLP Lab, Soochow University, Suzhou, P. R. China
shaohuikuang@foxmail.com

Abstract. In neural machine translation (NMT), researchers face the challenge of un-seen (or out-of-vocabulary OOV) words translation. To solve this, some researchers propose the splitting of western languages such as English and German into sub-words or compounds. In this paper, we try to address this OOV issue and improve the NMT adequacy with a harder language Chinese whose characters are even more sophisticated in composition. We integrate the Chinese radicals into the NMT model with different settings to address the unseen words challenge in Chinese to English translation. On the other hand, this also can be considered as semantic part of the MT system since the Chinese radicals usually carry the essential meaning of the words they are constructed in. Meaningful radicals and new characters can be integrated into the NMT systems with our models. We use an attention-based NMT system as a strong baseline system. The experiments on standard Chinese-to-English NIST translation shared task data 2006 and 2008 show that our designed models outperform the baseline model in a wide range of state-of-the-art evaluation metrics including LEPOR, BEER, and CharacTER, in addition to BLEU and NIST scores, especially on the adequacy-level translation. We also have some interesting findings from the results of our various experiment settings about the performance of words and characters in Chinese NMT, which is different with other languages. For instance, the fully character level NMT may perform well or the state of the art in some other languages as researchers demonstrated recently, however, in the Chinese NMT model, word boundary knowledge is important for the model learning. [3]

Keywords: Machine Translation · Chinese-English Translation · Chinese Radicals · Neural Networks · Translation Evaluation.

1 Introduction

Neural Machine Translation (NMT) models treat MT task as encoder-decoder work-flow which is much different from the conventional SMT structure [7]. The

[3] * parallel authors, ranked by alphabet order

Fig. 1: Radical as independent character.

encoder applies in the source language side learning the sentences into vector representations, while the decoder applies in the target language side generating the words from the target side vectors. Recurrent Neural Networks (RNN) models are usually used for both encoder and decoder, though there are some researchers employing convolutions neural networks (CNN) like [6, 15]. The hidden layers in the neural nets are designed to learn and transfer the information [22]. There were some drawbacks in the NMT models e.g. lack of alignment information between source and target side, and less transparency, etc. To address these, attention mechanism was introduced to the decoder first by [1] to pay interests to part information of the source sentence selectively, instead of the whole sentence always, when the model is doing translation.

Another drawback of NMT is that the NMT systems usually produce better fluent output, however, the adequacy is lower sometimes compared with the conventional SMT, e.g. some meaning from the source sentences will be lost in the translation side when the sentence is long [28, 29, 16, 22, 6]. One kind of reason of this phenomenon could be due to the unseen words problem, except for the un-clear learning procedure of the neural nets. With this assumption, we try to address the unseen words or out-of-vocabulary (OOV) words issue and improve the adequacy level by exploring the Chinese radicals into NMT.

For Chinese radical knowledge, let's see two examples about their construction in the corresponding characters. This Figure 1 shows three Chinese characters (forest, tree, bridge) which contain the same part of radical (wood) and this radical can be a character independently in usage. In the history, Chinese bridge was built by wood usually, so apparently, these three characters carry the similar meaning that they all contain something related with woods.

The Figure 2 shows three Chinese characters (grass, medicine, tea) which contain the same part of radical (grass) however this radical can not be a character independently in usage. This radical means grass in the original development of Chinese language. In the history, Chinese medicine was usually developed from some nature things like the grass, and Chinese tea was usually from the leafs that are related with grass. To the best knowledge of the authors at the submission stage, there is no published work about radical level NMT for Chinese language yet.

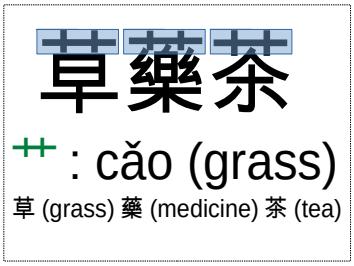


Fig. 2: Radical as non-independent character.

2 Related Work

MT models have been developed by utilizing smaller units, i.e. phrase-level to word-level, sub-word level and character-level [24, 8]. However, for Chinese language, sub-character level or radical level is also a quite interesting topic since the Chinese radicals carry somehow essential meanings of the Chinese characters that they are constructed in. Some of the radicals splitted from the characters can be independent new characters, meanwhile, there are some other radicals that can not be independent as characters though they also have meanings. It would be very interesting to see how these radicals or the combination of them and traditional words/characters perform in the NMT systems.

There are some published works about the investigation of Chinese radicals embedding for other tasks of NLP, such as [25, 18] explored the radical usage for word segmentation and text categorization.

Some MT researchers explored the word composition knowledge into the systems, especially on the western languages. For instance, [21] developed a Machine Translation model on English-German and English-Finnish with the consideration of synthesizing compound words. This kind of knowledge is similar like the splitting Chinese character into new characters.

3 Model Design

3.1 Attention-based NMT

Typically, as mentioned before, neural machine translation (NMT) builds on an encoder-decoder framework [1, 27] based on recurrent neural networks (RNN). In this paper, we take the NMT architecture proposed by [1]. In NMT system, the encoder apples a bidirectional RNN to encode a source sentence $x = (x_1, x_2, ..., x_{T_x})$ and repeatedly generates the hidden vectors $h = (h_1, h_2, ..., h_{T_x})$ over the source sentence, where T_x is the length of source sentence. Formally, $h_j = [\overrightarrow{h_j}; \overleftarrow{h_j}]$ is the concatenation of forward RNN hidden state $\overrightarrow{h_j}$ and backward RNN hidden state $\overleftarrow{h_j}$, and $\overrightarrow{h_j}$ can be computed as follows:

$$\overrightarrow{h_j} = f(\overrightarrow{h_{j-1}}, x_j) \tag{1}$$

where function f is defined as a Gated Recurrent Unit (GRU) [9].

The decoder is also an RNN that predicts the next word y_t given the context vector c_t, the hidden state of the decoder s_t and the previous predicted word y_{t-1}, which is computed by:

$$p(y_t|y_{<t}, x) = softmax(g(s_t, y_{t-1}, c_t)) \tag{2}$$

where g is a non-linear function. and s_t is the state of decoder RNN at time step t, which is calculated by:

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{3}$$

where c_t is the context represent vector of source sentence.

Usually c_t can be obtained by attention model and calculated as follows:

$$c_t = \sum_{j=1}^{T_x} \alpha_{tj} h_j \tag{4}$$

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{T_x} e_{tk}} \tag{5}$$

$$e_{tj} = v_a^T tanh(s_{t-1}, h_j) \tag{6}$$

We also follow the implementation of attention-based NMT of dl4mt tutorial [4], which enhances the attention model by feeding the previous word y_{t-1} to it, therefore the e_{tj} is calculated by:

$$e_{tj} = v_a^T tanh(\widetilde{s}_{t-1}, h_j) \tag{7}$$

where $\widetilde{s}_{t-1} = f(s_{t-1}, y_{t-1})$, and f is a GRU function. The hidden state of the decoder is updated as following:

$$s_t = f(\widetilde{s}_{t-1}, c_t) \tag{8}$$

In this paper, we use the attention-based NMT with the changes from dl4mt tutorial [5] as our baseline and call it RNNSearch*[6].

[4] github.com/nyu-dl/dl4mt-tutorial/tree/master/ session2

[5] github.com/nyu-dl/dl4mt-tutorial

[6] To distinguish it from RNNSearch as in the paper [1]

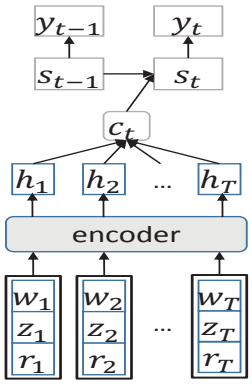


Fig. 3: Architecture of NMT with multi-embedding.

3.2 Our model

Traditional NMT model usually uses the word-level or character-level information as the inputs of encoder, which ignores some knowledge of the source sentence, especially for Chinese language. Chinese words are usually composed of multiple characters, and characters can be further splitted into radicals. The Chinese character construction is very complected, varying from upper-lower structure, left-right structure, to inside-outside structure and the combination of them. In this paper, we use the radical, character and word as multiple inputs of NMT and expect NMT model can learn more useful features based on the different levels of input integration.

Figure 3 illustrates our proposed model. The input embedding x_j consists of three parts: word embedding w_j, character embedding z_j [7] and radical embedding r_j, as follows:

$$x_j = [w_j; z_j; r_j] \tag{9}$$

where ';' is concatenate operation.

For the word w_j, it can be split into characters $z_j = (z_{j1}, z_{j2}, ..., z_{jm})$ and further split into radicals $r_j = (r_{j1}, r_{j2}, ..., r_{jn})$. In our model, we use simple additions operation to get the character representation and radical representation of the word, i.e. z_j and r_j can be computed as follows:

$$z_j = \sum_{k=1}^{m} z_{jk} \tag{10}$$

$$r_j = \sum_{k=1}^{n} r_{jk} \tag{11}$$

[7] We use the character 'z' to represent character, instead of 'c', because we already used 'c' as representation of context vector.

Each word can be decomposed into different numbers of character and radical, and, by addition operations, we can generate a fixed length representation. In principle our model can handle different levels of input from their combinations. For Chinese character decomposition, e.g. the radicals generation, we use the HanziJS open source toolkit [8]. On the usage of target vocabulary [14], we choose 30,000 as the volume size.

4 Experiments

4.1 Experiments Setting

We used 1.25 million parallel Chinese-English sentences for training, which contain 80.9 millions Chinese words and 86.4 millions English words. The data is mainly from Linguistic Data Consortium (LDC) [9] parallel corpora, such as LDC2002E18, LDC2003E07, LDC2003E14, LDC2004T07, LDC2004T08, and LDC2005T06. We tune the models with NIST06 as development data using BLEU metric [23], and use NIST08 Chinese-English parallel corpus as testing data with four references.

For the baseline model RNNSearch*, in order to effectively train the model, we limit the maximum sentence length on both source and target side to 50. We also limit both the source and target vocabularies to the most frequent 30k words and replace rare words with a special token "UNK" in Chinese and English. The vocabularies cover approximately 97.7% and 99.3% of the two corpora, respectively. Both the encoder and decoder of RNNsearch* have 1000 hidden units. The encoder of RNNsearch consists of a forward (1000 hidden unit) and backward bidirectional RNN. The word embedding dimension is set as 620. We incorporate dropout [13] strategy on the output layer. We used the stochastic descent algorithm with mini-batch and Adadelta [31] to train the model. The parameters ρ and ϵ of Adadelta are set to 0.95 and 10^{-6}. Once the RNNsearch* model is trained, we adopt a beam search to find possible translations with high probabilities. We set the beam width of RNNsearch* to 10. The model parameters are selected according to the maximum BLEU score points on the development set.

For our proposed model, all the experimental settings are the same as RNNSearch*, except for the word-embedding dimension and the size of the vocabularies. In our model, we set the word, character and radical to have the same dimension, all 620. The vocabulary sizes of word, character and radical are set to 30k, 2.5k and 1k respectively.

To integrate the character radicals into NMT system, we designed several different settings as demonstrated in the table. Both the baseline and our settings used the attention-based NMT structure.

[8] github.com/nieldlr/Hanzi
[9] www.ldc.upenn.edu

Settings	Description	abbreviation
Baseline	Words	W
Setting1	Word+Character+Radical	W+C+R
Setting2	Word+Character	W+C
Setting3	Word+Radical	W+R
Setting4	Character+Radical	C+R

Table 1: Model Settings

4.2 Evaluations

Firstly, there are many works reflecting the insufficiency of BLEU metric, such as higher or lower BLEU scores do not necessarily reflect the model quality improvements or decreasing; BLEU scores are not interpretive by many translation professionals; and BLEU did not correlate better than later developed metrics in some language pairs [5, 4, 17].

In the light of such analytic works, we try to validate our work in a deeper and broader evaluation setting from more aspects. We use a wide range of state of the art MT evaluation metrics, which are developed in recent years, to do a more comprehensive evaluation, including hLEPOR [11, 12], CharacTER [30], BEER [26], in addition to BLEU and NIST [23].

The model hLEPOR is a tunable translation evaluation metric yielding higher correlation with human judgments by adding n-gram position difference penalty factor into the traditional F-measures. CharacTER is a character level editing distance rate metric. BEER uses permutation trees and character n-grams integrating many features such as paraphrase and syntax. They have shown top performances in recent years' WMT[10] shared tasks [20, 19, 10, 3].

Both CharacTER and BEER metrics achieved the parallel top performance in correlation scores with human judgment on Chinese-to-English MT evaluation in WMT-17 shared tasks [2] . While LEPOR metric series are evaluated by MT researchers as one of the most distinguished metric families that are not apparently outperformed by others, which is stated in the metrics comparison work in [10] on standard WMT data.

Evaluation on Development Set On the development set NIST06, we got the following evaluation scores. The cumulative N-gram scoring of BLEU and NIST metric, with bold case as the highlight of the winner in each n-gram column situation, is shown in the table respectively. Researchers usually report their 4-gram BLEU while 5-gram NIST metric scores, so we also follow this tradition here:

From the scoring results, we can see that the model setting one, i.e. W+C+R, won the baseline models in all uni-gram to 4-gram BLEU and to 5-gram NIST scores. Furthermore, we can see that, by adding character and/or radical to the words, the model setting two and three also outperformed the baseline models.

[10] www.statmt.org/wmt17/metrics-task.html

	1-gram	2-gram	3-gram	4-gram
Baseline	.7211	.5663	.4480	.3556
W+C+R	**.7420**	**.5783**	**.4534**	**.3562**
W+C	.7362	.5762	.4524	.3555
W+R	.7346	.5730	.4491	.3529
C+R	.7089	.5415	.4164	.3219

Table 2: BLEU Scores on NIST06 Development Data

	1-gram	2-gram	3-gram	4-gram	5-gram
Baseline	5.8467	7.7916	8.3381	8.4796	8.5289
W+C+R	**6.0047**	**7.9942**	**8.5473**	**8.6875**	**8.7346**
W+C	5.9531	7.9438	8.5127	8.6526	8.6984
W+R	5.9372	7.9021	8.4573	8.5950	8.6432
C+R	5.6385	7.4379	7.9401	8.0662	8.1082

Table 3: NIST Scores on NIST06 Development Data

However, the setting 4 that only used character and radical information in the model lost both BLEU and NIST scores compared with the word-level baseline. This means that, for Chinese NMT, the word segmentation knowledge is important to show some guiding in Chinese translation model learning.

For uni-gram BLEU score, our Model one gets 2.1 higher score than the baseline model which means by combining W+C+R the model can yield higher adequacy level translation, though the fluency score (4-gram) does not have much difference. This is exactly the point that we want to improve about neural models, as complained by many researchers.

The evaluation scores with broader state-of-the-art metrics are shown in the following table. Since CharacTER is an edit distance based metric, the lower score means better translation result.

	Metrics on Single Reference		
Models	hLEPOR	BEER	CharacTER
Baseline	.5890	.5112	.9225
W+C+R	.5972	**.5167**	**.9169**
W+C	**.5988**	.5164	.9779
W+R	.5942	.5146	.9568
C+R	.5779	.4998	1.336

Table 4: Broader Metrics Scores on NIST06 Development Data

From the broader evaluation metrics, we can see that our designed models also won the baseline system in all the metrics. Our model setting one, i.e. the W+C+R model, won both BEER and CharacTER scores, while our model two, i.e. the W+C, won the hLEPOR metric score, though the setting four continue

to be the worest performance, which is consistent with the BLEU and NIST metrics. Interestingly, we find that the CharacTER score of setting two and three are both worse than the baseline, which means that by adding of character and radical information separately the output translation needs more editing effort; however, if we add both the character and radical information into the model, i.e. the setting one, then the editing effort became less than the baseline.

Evaluation on Test Sets The evaluation results on the NIST08 Chinese-to-English test date are presented in this section.

Firstly, we show the evaluation scores on BLEU and NIST metrics, with four reference translations and case-insensitive setting. The tables show the cumulative N-gram scores of BLEU and NIST, with bold case as the winner of each n-gram situation in each column.

	1-gram	2-gram	3-gram	4-gram
Baseline	.6451	.4732	.3508	.2630
W+C+R	**.6609**	**.4839**	**.3572**	**.2655**
W+C	.6391	.4663	.3412	.2527
W+R	.6474	.4736	.3503	.2607
C+R	.6378	.4573	.3296	.2410

Table 5: BLEU Scores on NIST08 Test Data

	1-gram	2-gram	3-gram	4-gram	5-gram
Baseline	5.1288	6.6648	7.0387	7.1149	7.1387
W+C+R	**5.2858**	**6.8689**	**7.2520**	**7.3308**	**7.3535**
W+C	5.0850	6.5977	6.9552	7.0250	7.0467
W+R	5.1122	6.6509	7.0289	7.1062	7.1291
C+R	5.0140	6.4731	6.8187	6.8873	6.9063

Table 6: NIST Scores on NIST08 Test Data

The results show that our model setting one won both BLEU and NIST scores on each n-gram evaluation scheme, while model setting three, i.e. the W+R model, won the uni-gram and bi-gram BLEU scores, and got very closed score with the baseline model in NIST metric. Furthermore, the model setting four, i.e. the C+R one, continue showing the worst ranking, which may verify that word segmentation information and word boundaries are indeed helpful to Chinese translation models, so we can not omit such part.

What worth to mention is that the detailed evaluation scores from BLEU reflect our Model one yields higher BLEU score (1.58) on uni-gram, similar with the results on development data, while a little bit higher performance on 4-gram

(0.25). These mean that in the fluency level our translation is similar with the state-of-the-art baseline, however, our model yields much better adequacy level translation in NMT since uni-gram BLEU reflects the adequacy aspect instead of fluency. This verifies the value of our model in the original problem we want to address.

The evaluation results on recent years' advanced metrics are shown below. The scores are also evaluated on the four references scheme. We calculate the average score of each metric from 4 references as the final evaluation score. Bold case means the winner as usual.

	Metrics Evaluated on 4-references		
Models	hLEPOR	BEER	CharacTER
Baseline	.5519	.4748	**0.9846**
W+C+R	**.5530**	**.4778**	1.3514
W+C	.5444	.4712	1.1416
W+R	.5458	.4717	0.9882
C+R	.5353	.4634	1.1888

Table 7: Broader Metrics Scores on NIST08 Test Data

From the broader evaluations, we can see that our model setting one won both the LEPOR and BEER metrics. Though the baseline model won the CharacTER metric, the margin between the two scores from baseline (.9846) and our model three, i.e. W+R, (.9882) is quite small around 0.0036. Continuously, the setting four with C+R performed the worst though and verified our previous findings.

5 Conclusion and Future Work

We presented the different performances of the multiple model settings by integrating Chinese character and radicals into state-of-the-art attention-based neural machine translation systems, which can be helpful information for other researchers to look inside and gain general clues about how the radical works.

Our model shows the full character+radical is not enough or suitable for Chinese language translation, which is different with the work on western languages such as [8]. Our model results showed that the word segmentation and word boundary are helpful knowledge for Chinese translation systems.

Even though our model settings won both the traditional BLEU and NIST metrics, the recent years developed advanced metrics indeed showed some differences and interesting phenomena, especially the character level translation error rate metric CharacTER. This can encourage MT researchers to use the state-of-the-art metrics to find useful insight of their models.

Although the combination of words, characters and radicals mostly yielded the best scores, the broad evaluations also showed that the model setting W+R, i.e. using both words and radicals information, is generally better than the model

setting W+C, i.e. words plus characters without radical, which verified the value of our work by exploring radicals into Chinese NMT. Our Model one yielded much better adequacy level translation output (by uni-gram BLEU score) compared with the baseline system, which also showed that this work is important in exploring how to improve adequacy aspect of neural models.

In the future work, we will continue to optimize our models and use more testing data to verify the performances. In this work, we aimed at exploring the effectiveness of Chinese radicals, so we did not use BPE for English side splitting, however, to promote the state-of-the-art Chinese-English translation, in our future extension, we will apply the splitting on both Chinese and English sides. We will also investigate the usage of Chinese radicals into MT evaluation area, since they carry the language meanings.

6 Acknowledgement

The author Han thanks Ahmed Abdelkader for the kind help, and Niel de la Rouviere for the HanziJS toolkit. This work was supported by Soochow University of China and ADAPT Centre of Ireland. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR **abs/1409.0473** (2014), `http://arxiv.org/abs/1409.0473`
2. Bojar, O., Graham, Y., Kamran, A.: Results of the WMT17 metrics shared task. In: Proceedings of WMT, Vol 2: Shared Tasks Papers. Association for Computational Linguistics, Copenhagen, Denmark (September 2017)
3. Bojar, O., Graham, Y., Kamran, A., Stanojević, M.: Results of the wmt16 metrics shared task. In: WMT. pp. 199–231 (2016)
4. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (meta-) evaluation of machine translation. In: WMT. pp. 64–71 (2007)
5. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of bleu in machine translation research. In: Proceedings of EACL. vol. 2006, pp. 249–256 (2006)
6. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. CoRR **abs/1409.1259** (2014)
7. Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: EMNLP (2014)
8. Chung, J., Cho, K., Bengio, Y.: A character-level decoder without explicit segmentation for neural machine translation. In: ACL (2016)
9. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS Deep Learning and Representation Learning Workshop (2014)

10. Graham, Y., Mathur, N., Baldwin, T.: Accurate evaluation of segment-level machine translation metrics. In: Proceedings of NAACL-HLT. Denver, Colorado (2015)
11. Han, A.L.F., Wong, D.F., Chao, L.S., He, L., Lu, Y., Xing, J., Zeng, X.: Language-independent model for machine translation evaluation with reinforced factors. In: Machine Translation Summit XIV. pp. 215–222. IAMT (2013)
12. Han, L.: LEPOR: An Augmented Machine Translation Evaluation Metric. University of Macau, Thesis (2014), http://arxiv.org/abs/1703.08748
13. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
14. Jean, S., Cho, K., Memisevic, R., Bengio, Y.: On using very large target vocabulary for neural machine translation. In: ACL 2015 (2014)
15. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. Association for Computational Linguistics, Seattle (October 2013)
16. Koehn, P., Knowles, R.: Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872 (2017)
17. Lavie, A.: Automated metrics for mt evaluation. Machine Translation **11**, 731 (2013)
18. Liu, F., Lu, H., Lo, C., Neubig, G.: Learning character-level compositionality with visual features. CoRR **abs/1704.04859** (2017), http://arxiv.org/abs/1704.04859
19. Machacek, M., Bojar, O.: Results of the wmt14 metrics shared task. In: WMT. pp. 293–301 (2014)
20. Macháček, M., Bojar, O.: Results of the WMT13 metrics shared task. In: WMT. pp. 45–51 (2013)
21. Matthews, A., Schlinger, E., Lavie, A., Dyer, C.: Synthesizing compound words for machine translation. In: ACL (1) (2016)
22. Neubig, G.: Neural machine translation and sequence-to-sequence models: A tutorial. arXiv preprint arXiv:1703.01619 (2017)
23. Papineni, K., Roukos, S., Ward, T., jing Zhu, W.: Bleu: a method for automatic evaluation of machine translation. pp. 311–318 (2002)
24. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. CoRR **abs/1508.07909** (2015), http://arxiv.org/abs/1508.07909
25. Shi, X., Zhai, J., Yang, X., Xie, Z., Liu, C.: Radical embedding: Delving deeper to chinese radicals. In: Proceedings of ACL-IJCNLP. pp. 594–598. Association for Computational Linguistics (2015)
26. Stanojević, M., Sima'an, K.: Beer: Better evaluation as ranking. In: Proceedings of WMT. ACL (2014)
27. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
28. Tu, Z., Liu, Y., Lu, Z., Liu, X., Li, H.: Context gates for neural machine translation. CoRR **abs/1608.06043** (2016), http://arxiv.org/abs/1608.06043
29. Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H.: Coverage-based neural machine translation. CoRR **abs/1601.04811** (2016), http://arxiv.org/abs/1601.04811
30. Wang, W., Peter, J.T., Rosendahl, H., Ney, H.: Character: Translation edit rate on character level. In: WMT. pp. 505–510 (2016)
31. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)

Towards a Cognitive Model of the Semantics of Spatial Prepositions

Adam Richard-Bollans

University of Leeds, Leeds, UK
mm15alrb@leeds.ac.uk

Abstract. There has been much work in linguistics and cognitive science regarding the semantics of spatial prepositions, highlighting the complexity of representing their meaning. However, much of this insight has not yet been translated into a satisfactory computational model. A central problem with existing frameworks is the lack of non-geometric features. In this paper we firstly argue for *cognitive adequacy* of the semantic representation and we consider how conceptual spaces may aid this representation task. We then consider how salient features, both geometric and non-geometric, may be integrated into the framework such that the model is more closely aligned with the way humans conceptualize spatial prepositions.

Keywords: Semantics · Spatial cognition · Conceptual spaces

1 Introduction

The aim of this project is to provide a robust, cognitively-aligned framework for modelling spatial prepositions that can be used for natural language understanding and generation. In particular we consider (1) the problem of selecting appropriate objects when given locative descriptions and (2) the problem of generating appropriate locative referring expressions.

There are a relatively small number of spatial prepositions in the English language which are used to encode a potentially infinite set of possible configurations of entities. This necessitates spatial prepositions to be broad and flexible in their meaning, exhibiting vagueness and polysemy, and poses problems for many systems where commands or queries are given in natural language.

There has been much work in linguistics and cognitive science regarding the meaning of spatial prepositions, highlighting the different roles various features can play in influencing preposition use. However, this insight has not yet been translated into a satisfactory computational model. Existing computational models of spatial prepositions for robots/visual systems tend to be on very constrained environments or give preference to big data over linguistic understanding.

As a preliminary task to assess the adequacy of our model we intend to construct virtual table top environments containing many similar objects in various configurations. These environments will also incorporate physics engines in order

to assess features such as 'the degree of location control'. Human subjects will firstly be asked to annotate the environments by assigning spatial prepositions to pairs of objects. Subjects will also be asked to identify specific objects when given spatial definite descriptions e.g. 'the pencil in the mug'. We will then use this data to help construct and test our framework.

1.1 Motivation

The main motivation for this project is to help mediate human-robot interaction. Humans often prefer brief, ambiguous descriptions over lengthy, unambiguous descriptions [41] and locative expressions often fulfil this desire for brevity. For example, rather than referring to objects based on elaborate visual attributes like 'the yellow cup with two pink dots on it', humans often refer to objects using simple locative expressions, say 'the cup next to the stapler'. We also see many examples of these expressions in the SemEval-2014 corpus [16] and the HuRIC corpus [6], both of which consider natural language commands given to robots.

2 Related Work

In this section we firstly give a brief description of some attempts to model the semantics of spatial prepositions, in particular those creating representations in a geometric space. We then consider more practical implementations.

2.1 Geometric Models

Though the main focus of cognitive modelling using geometric spaces has often been on non-relational concepts such as natural kinds [39], there has been more recent work focussing on relational concepts.

We first consider the early work of Abella and Kender [1] who aim to represent the semantics of spatial prepositions in a multidimensional geometric space. They define the prepositions *near, far, inside, above, below, aligned* and *next* based on twelve physical properties such as object area and shortest distance between two objects. These properties form the dimensions of the space. Each preposition is then represented as a set of inequalities along each dimension which creates a set of points in the given 12-D space. Though their work is preliminary, Abella and Kender hope that by using a spatial representation new insights will be gained into the prepositions and how they relate to each other. Due to the inherent vagueness of spatial prepositions Abella and Kender argue for 'fuzzification' to be incorporated in the model. This fuzzification is to be achieved using *fuzzy sets* [46], beginning with 'ideal' regions for prepositions, this is a crisp set, and then a fuzzy set is given by a distance function from the ideal region.

Since the influential work of Gärdenfors on conceptual spaces [24] there has been an increased interest in geometric semantic representations, see [13, 42, 49]. The geometric structure that makes up a conceptual space is a one or multi-dimensional metric space where the dimensions of the space, known as 'quality

dimensions', represent qualities of concepts such as temperature, weight, brightness etc.. Some quality dimensions are grouped as 'domains', for example the colour domain is composed of the hue, saturation and brightness dimensions.

The work of Gärdenfors on spatial prepositions [25] provides strong motivation for the direction taken in the current paper, so we outline this work here. The central thesis of his paper is that prepositions can be represented by convex[1] sets in a single domain. Gärdenfors considers two classes of prepositions, those based on the spatial domain and those based on the force domain. For those based on the spatial domain Gärdenfors provides various representations using a notion of 'betweenness' based on polar co-ordinates such that the representing sets are convex. Gärdenfors then posits that 'in', 'on' and 'against' belong to the force domain but does not provide an explicit method for representing them.

We however diverge from this thesis as, though we agree that some spatial prepositions may be based on a single domain, we believe that it is necessary to take into account various aspects of meaning in the representation. The spatial and force domains are intimately linked, which Gärdenfors does also concede in stating that, for example, "there are examples where "in" is used purely spatially". See Sections 3.2 & 4 for further discussion on this issue.

2.2 Implementations

Here we consider related implementations, what we can learn from them and how they may be improved.

In a task focussed on discourse and interaction [32] Moratz and Tenbrink consider real world scenarios in which human subjects give locative descriptions of objects to a robot which then has to move to the correct 'goal' object. They create a computational model of projective prepositions, incorporating intrinsic and relative reference frames and then experiment in a real environment.

Moratz and Tenbrink take an iterative approach, experimenting, analysing and improving their system. This type of iterative approach recognizes the complexity of discourse and allowed the researchers to identify strategies adopted by the speakers in order to improve the system. We intend to emulate this iterative approach in our implementations. From the experiments carried out they were able to recognize various tendencies of the human subjects, for example 'speakers intuitively use the robot's perspective' and 'applicability regions for spatial expressions seem to be fairly large; they are not mutually exclusive' [32].

So far, we have only considered hard-coded models of spatial prepositions though there are many instances of trained models for similar tasks [3, 22, 23].

Guadarrama et al. [23] consider a task similar to [32], where objects are set out on a table top and a robot must select the correct figure object[2] when given

[1] A region R is *convex* if for any two points x, y in R, if z is *between* x and y then z is also in R. Note that various metrics can be assigned which alter the notion of *between*

[2] The *figure object* (also known as: target, trajector, referent) is the entity whose location is important e.g. 'the **bike** next to the house'

a ground object[3] and spatial preposition. They consider that trained models of spatial prepositions perform better than hard-coded ones and so they use a statistical model which is trained using annotated 3-D GoogleSketchup[4] models. They consider various models of features (simple, complex, psycholinguistic) which allow for the creation of a hybrid model. The simple and complex features are similar to that of [1] while the psycholinguistic features come from the more cognitively motivated discussion of [36].

The recent work of Alomari et al. [3], though heavily focused on machine learning, contains similar ideas to the current project; in particular with respect to concept formation. Their work considered a robot with a table top environment where the robot had to perform various manipulation tasks based on natural language commands such as 'place the apple in the bowl'. The robot was able to successfully ground the language by training on segmented videos paired with corresponding natural language commands as well as using a surprisingly small amount of hard-coded knowledge.

Though machine learning and big data techniques for such commonsense tasks can be attractive, as we discuss in [38], an over-reliance on statistical methods can be problematic when dealing with the intricacies of natural language. It may be the case that with enough training data one of these statistical models creates an internal representation that is closely aligned with a satisfactory cognitive model. However, such models are likely to be highly context sensitive [14, 33], uninterpretable by humans and difficult to update on-the-fly. Also, there is a lack of corpora for training for this task [5, 6].

The current project will be closely aligned with a lot of the themes present in the pieces of work above, though as a fundamental difference we will explicitly incorporate functional and commonsense aspects of preposition use.

3 Modelling Spatial Prepositions

There has been much debate on what can be meaningfully said about spatial prepositions and how they can be appropriately modelled [7, 12, 26, 47]. From this we see a clear general consensus that to provide an adequate account of the semantics of spatial prepositions non-geometric features must be considered.

There have been numerous attempts to model spatial prepositions [22, 28, 30, 40]. However, these methods do not yet incorporate significant non-geometric salient features.

3.1 Cognition, Discourse & Conceptual Spaces

We believe that to appropriately pin down the semantics of natural language one must use cognition as a basic point of reference — an ideal semantic representation should be *cognitively adequate* [44]. Cognitive adequacy aids in understanding the conceptual representation employed by the people we are speaking

[3] The *ground object* (also known as: reference, landmark, relatum) is the entity used as a reference point in order to locate the figure e.g. 'the bike next to the **house**'

[4] http://sketchup.google.com/3dwarehouse/

to. We therefore believe that ideal semantic models for this task should be akin to human conceptual models.

Relating to cognition and processing visual scenes, Herskovits [27] proposes that humans generate locative expressions via *schematization*. Schematization is the process of abstracting a physical scene from rich perceptual data to a sparse semantic representation via idealization and approximation. In order to adequately capture the meaning of spatial prepositions we must therefore consider the mapping between language and the perceptual system [12].

Further, we recognize that, particularly during situated discourse, meaning is not fixed across domains and individuals; nor even do individuals hold permanent interpretations of the meaning of terms [17]. Context and individual preference strongly influence how words are used and how their meaning is understood. Therefore, attaining an understanding of an interlocutor's conceptual representation is often a process that continues throughout discourse; the meaning of terms emerge and are refined through conversation and context. We therefore desire a semantic model which can be updated in a natural way throughout discourse.

We propose that the influential cognitive framework of conceptual spaces [24] provides a fitting semantic framework for this task. Conceptual spaces are cognitively motivated, aim to make explicit the link between perceptual information and symbolic expressions and provide potential mechanisms for naturally updating concepts.

Regarding the ability to naturally update concepts, regions in a conceptual space are ideally convex. Therefore, if it is necessary to generalize a concept by including a new point or region, we can update the concept via taking the convex hull[5] of the new points with the old region. This hopefully provides a felicitous way for generalizing concepts. This proposal requires further investigation, however, as of course this would not help when refining a concept and also depends on the ability to achieve convex representations.

This section has provided motivation for the use of conceptual spaces. In Section 4, after further discussion on the semantics of spatial prepositions, we outline more precise details of which features should be included in the conceptual space.

3.2 Non-Geometric Aspects of Spatial Preposition Meaning

Distinguishing between the uses of spatial prepositions requires a consideration of more than just relative locations of objects. This is a widely accepted view in linguistics and cognitive science [12, 20, 21, 26, 48]. For example, the curvature of a ground object influences the choice of 'in' or 'on' — explaining the use of 'in' for bowls and 'on' for plates [20]. Further, it is well accepted that the force dynamics of objects heavily influences preposition use and that this should be included in semantic models [12].

[5] The convex hull of a set of points, X, is the smallest convex set containing X

Some salient features, such as distance between two objects or their functional interaction, may be identifiable through perception of a scene. We call these features *contextual*. However, there are also many non-contextual features that influence preposition use such as affordances of objects and convention [12, 26]. We call these features *inherent*.

Identifying the inherent features requires commonsense knowledge of properties of objects and how they relate to each other. Though recognized as essential for a full understanding of preposition use, the incorporation of commonsense knowledge is an unexplored avenue in attempts to model spatial prepositions. Indeed, most related work considers indistinct objects such as square blocks that are not associated with particular conventions.

Tentative proposals for which features to include and how they can be incorporated into a conceptual space are discussed in the following section.

4 Salient Features

In order to provide more detail regarding salient features and their integration into a conceptual space, we outline some important categories of features and begin to construct a taxonomy.

4.1 Taxonomy of Features

We believe that the majority of features discussed in the literature can be split into the following three broad categories :

1. Geometric: Relating to spatial aspects of a given scene e.g. the shortest distance between two objects
2. Functional: Relating to the functional interplay between objects in a scene e.g. the extent to which one object supports another or if there is a functional electric connection between two objects
3. Conventional: Relating to how terms are commonly used e.g. one usually says 'on the bus' but 'in the car' even though the geometric and functional relations are very similar

These broad categories provide a basis for most spatial prepositions. In this paper we further restrict attention to the particular case of 'in', see Figure 1.

We highlight two particularly salient geometric features for 'in' as 'containment' and 'contact'. Next, we consider functional aspects of 'in' — the extent to which the ground object exerts location control over the figure object and the 'usual' degree of location control that this type of ground object can assert. Finally we consider typical usage of prepositions in order to account for conventional distinctions e.g. '*in* the bus' or '*on* the bus'. The process of assessing and refining these features is ongoing.

As shown in Figure 1 we recognize that there is significant interplay between these categories e.g. geometric features can directly imply functional ones, conventional features can imply functional ones etc..

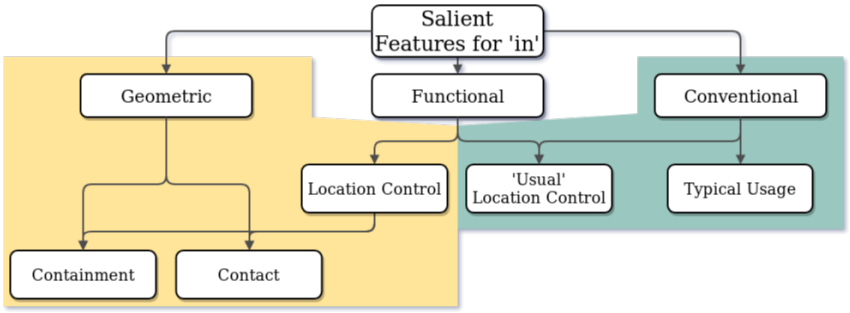


Fig. 1. The yellow section represents contextual features while teal represents inherent features

We now consider each of these features in turn and how they can be appropriately categorized, providing some motivation for the distinctions drawn and descriptions of how they may be valued. At present, due to the significant research challenge that it presents (see [2, 11, 18, 19, 34, 45]), we omit some issues relating to the extraction of qualitative spatial descriptions from visual scenes.

Containment The notion of containment is often cited as fundamental to the preposition 'in' [7, 26]. Containment, however, arises in many different forms depending on how the objects and the environments are idealized, consider the usage of 'in' for 'in a cup', 'crack in a glass' and 'in the town' [7].

Considering the work of Cohn et al. [10], we can see that the Region Connection Calculus [35] with the convex hull operator is expressive enough to distinguish various important types of containment. Some empirical support for the cognitive adequacy of the Region Connection Calculus is provided in [29, 31, 37].

Contact Though 'in' isn't immediately associated with contact, we believe that it can still play an important role. For example, it may be helpful for distinguishing 'location control via containment' and 'location control via contact'. Motivated by the distinctions drawn in results from experimental psychology [4], we intend to distinguish various types of contact e.g. 'on' and 'against'.

Degree of Location Control This feature is functionally motivated but (mostly) geometrically derived. With initial implementations we will determine this by running various simulations which assess whether a movement of the ground object moves the figure object. However, it may be that crude methods for determining location control, which translate better to the real world, will suffice and this is an avenue of further work. The inclusion of location control is motivated by the experiments of [21].

Usual Degree of Location Control There appears to be general consensus that 'in' rather than 'on' is appropriate in cases where the ground object could be described as a *container*. This is likely motivated by the location control that containers usually afford. However, the way humans conceptualize items affects preposition use [12]. When given exactly the same scene of an object on a plate/dish, humans will describe the configuration as *in* when the 'plate' is labelled as a dish, and *on* when labelled as a plate. For this reason, we consider it important to incorporate such information in the semantic model.

We propose achieving this via two stages. First we intend to identify certain attributes which are associated with location control such as 'concavity' or 'container'. Secondly, we intend to use commonsense knowledge in order to attain a value for how closely related the ground object is to these attributes. For example, this may be achieved by considering the weightings applied between concepts in ConceptNet [43].

We, however, recognize that this feature may be subsumed by the following conventional feature on typical word usage.

Typical Word Usage In order to further integrate conventional language use we should consider how common certain prepositions are with certain figure/ground objects. This may be achieved via simple metrics such as results from search engines or by combing through corpora. We make no reference to the cognitive adequacy of conventional features, though we believe their inclusion is clearly justified.

4.2 Integration within a Conceptual Space

We propose that each feature is used as a quality dimension where the functional and conventional dimensions have numerical values representing, for example, the *degree* of which the ground object exerts location control over the figure object. In these cases the standard Euclidean metric can be applied.

Some of the geometric quality dimensions will, however, take distinct geometric relations as values. We propose that containment, for example, be described by a set qualitative spatial relations. In order to represent this in a conceptual space we can apply a metric between these relations via a *conceptual neighbourhood graph*, as in [42].

5 Implementation

In initial implementations we intend to model 3D table-top environments. These models will also integrate physics engines, allowing for assessment of factors such as location control. This allows preliminary testing of the semantic model prior to a full-blown real-world implementation.

These models will be used to collect data and also to test the semantic representation, as described below.

5.1 Data Collection

Firstly, we highlight that there are distinct tasks that require distinct representations and/or methods of interpretation. In this project we consider the two broad categories of language generation and language interpretation. The generation side requires that the system can produce appropriate locative expressions to describe objects in a given scene. The interpretation side requires that the system can identify objects being referred to in a scene when given locative expressions.

We can see that there are slight differences in how the semantics are dealt with for these two tasks. For example, when describing somebody, Alan say, sat on a bus it would be inappropriate for the system to generate 'Alan is in a bus'. However, given that, say, Alan is sitting alone on a bus, it would be expected that the system interpret 'the person in the bus' as 'Alan'.

The data collected will reflect this distinction. We intend to have human subjects firstly annotate visual scenes describing relationships between objects using spatial prepositions. We will then ask subjects to identify objects in the scene when given definite descriptions e.g. 'the book on the chair' as well as answer questions such as 'is the book on the chair?'.

5.2 Concept Creation & Testing

Assuming that quality dimensions have been assigned, the data collected will be input into the conceptual space, giving a conceptual space populated by numerous points, or instances, of a given preposition. In order to do this, various features from the scene must be extracted and initially this will be done by a human expert. Acquisition of the geometric features by an expert is likely to be simple and accurate, however the precise details of evaluating functional and conventional features are still to be determined.

The resulting conceptual space must then be analysed to assess if the instances can be grouped in a meaningful way. In an ideal scenario this would be done by taking the convex hull of every point in the space. However, this is likely to create over-generalized representations. What we expect to find is that polysemes become apparent in the space. If this is the case we will need to cluster the points together. This is not a novel idea, see [9, 15]. Also, the conceptual space implementation of [8] supports sufficient operations for creation of such clustering algorithms.

The process of creating an appropriate conceptual space will be iterative — creating quality dimensions, inputting data and then using clustering to distinguish concepts. We will then analyse the clusters that arise to decide if they are meaningful and/or if they miss important examples.

6 Conclusion & Future Work

We believe that we can provide a framework for understanding and generating natural language along similar lines to Gärdenfors [25] which incorporates a

broader range of features. We intend to achieve this by drawing on the existing wealth of research in the topic, deepening the semantic analysis and adding quality dimensions in the conceptual space to reflect non-geometric salient features.

Once a representation space has been established we intend to analyse exactly how this can be employed in discourse and to further investigate the semantics of spatial prepositions. Firstly, we will consider how the representation can be updated to model an interlocutor's standpoint during discourse. Secondly, we will examine whether prototypes can be assigned in the space along with an appropriate metric. This will aid the process of disambiguating ambiguous definite descriptions. Finally, we will investigate whether the framework sheds light on the polysemous nature of spatial prepositions and if distinct polysemes can be identified.

Acknowledgements

Thanks to the anonymous reviewers for their detailed and insightful feedback and to Anthony Cohn and Brandon Bennett for helpful discussion on the topic.

References

1. Abella, A., Kender, J.R.: Qualitatively describing objects using spatial prepositions. In: IEEE Workshop on Qualitative Vision. pp. 33–38. IEEE (1993)
2. Albath, J., Leopold, J.L., Sabharwal, C.L., Maglia, A.M.: RCC-3d: Qualitative Spatial Reasoning in 3d. In: CAINE. pp. 74–79 (2010)
3. Alomari, M., Duckworth, P., Hogg, D.C., Cohn, A.G.: Natural Language Acquisition and Grounding for Embodied Robotic Systems. pp. 4349–4356 (2017)
4. Baillargeon, R.: The acquisition of physical knowledge in infancy: A summary in eight lessons. In: Usha, G. (ed.) Blackwell handbook of childhood cognitive development, pp. 47–83. Blackwell (2002)
5. Barclay, M., Galton, A.: A Scene Corpus for Training and Testing Spatial Communication Systems. In: AISB 2008 Convention Communication, Interaction and Social Intelligence. vol. 1, p. 26 (2008)
6. Bastianelli, E., Castellucci, G., Croce, D., Iocchi, L., Basili, R., Nardi, D.: HuRIC: a Human Robot Interaction Corpus. In: LREC. pp. 4519–4526 (2014)
7. Bateman, J.A., Hois, J., Ross, R., Tenbrink, T.: A linguistic ontology of space for natural language processing. Artificial Intelligence **174**(14), 1027–1071 (2010)
8. Bechberger, L., Kühnberger, K.U.: A Thorough Formalization of Conceptual Spaces. In: Kern-Isberner, G., Fürnkranz, J., Thimm, M. (eds.) KI 2017: Advances in Artificial Intelligence. pp. 58–71. Springer International Publishing, Cham (2017)
9. Cao, G., Song, D., Bruza, P.: Fuzzy K-means clustering on a high dimensional semantic space. In: Asia-Pacific Web Conference. pp. 907–911. Springer (2004)
10. Cohn, A.G., Bennett, B., Gooday, J., Gotts, N.M.: Qualitative spatial representation and reasoning with the region connection calculus. GeoInformatica **1**(3), 275–316 (1997)

11. Cohn, A.G., Magee, D.R., Galata, A., Hogg, D.C., Hazarika, S.M.: Towards an architecture for cognitive vision using qualitative spatio-temporal representations and abduction. In: International Conference on Spatial Cognition. pp. 232–248. Springer (2002)
12. Coventry, K.R., Garrod, S.C.: Saying, seeing and acting: The psychological semantics of spatial prepositions. Psychology Press (2004)
13. Cubek, R., Ertel, W.: Conceptual similarity as a key to high-level robot programming by demonstration. In: Proceedings of ROBOTIK 2012. pp. 1–6. VDE (2012)
14. Daumé III, H., Marcu, D.: Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research **26**, 101–126 (2006)
15. Douven, I.: Clustering colors. Cognitive Systems Research **45**, 70–81 (2017)
16. Dukes, K.: SemEval-2014 Task 6: Supervised Semantic Parsing of Robotic Spatial Commands. In: Proceedings of the 8th International Workshop on Semantic Evaluation. pp. 45–53 (2014)
17. Egré, P., de Gardelle, V., Ripley, D.: Vagueness and Order Effects in Color Categorization. Journal of Logic, Language and Information **22**(4), 391–420 (2013)
18. Falomir, Z.: Teaching Spatial Thinking, Computer Vision, and Qualitative Reasoning Methods. In: Proceedings of the Workshop on Teaching Spatial Thinking from Interdisciplinary Perspectives. pp. 11–15 (2015)
19. Falomir, Z., Kluth, T.: Qualitative spatial logic descriptors from 3d indoor scenes to generate explanations in natural language. Cognitive Processing pp. 1–20 (2017)
20. Feist, M.I., Gentner, D.: On plates, bowls, and dishes: Factors in the use of English IN and ON. In: Proceedings of the twentieth annual meeting of the cognitive science society. pp. 345–349 (1998)
21. Garrod, S., Ferrier, G., Campbell, S.: In and on: investigating the functional geometry of spatial prepositions. Cognition **72**(2), 167–189 (1999)
22. Golland, D.: Semantics and Pragmatics of Spatial Reference. PhD Thesis, University of California, Berkeley, USA (2013)
23. Guadarrama, S., Riano, L., Golland, D., Go, D., Jia, Y., Klein, D., Abbeel, P., Darrell, T.: Grounding spatial relations for human-robot interaction. pp. 1640–1647. IEEE (2013)
24. Gärdenfors, P.: Conceptual spaces: The geometry of thought. MIT press (2000)
25. Gärdenfors, P.: The Geometry Of Preposition Meanings. Baltic International Yearbook of Cognition, Logic and Communication **10**(1) (2015)
26. Herskovits, A.: Language and spatial cognition. Cambridge University Press (1987)
27. Herskovits, A.: Language, spatial cognition, and vision. In: Stock, O. (ed.) Spatial and Temporal Reasoning, pp. 155–202. Kluwer Academic Publishers, Dordrecht (1997)
28. Hois, J., Kutz, O.: Natural Language Meets Spatial Calculi. In: Freksa, C., Newcombe, N.S., Gärdenfors, P., Wölfl, S. (eds.) Spatial Cognition VI. Learning, Reasoning, and Talking about Space, vol. 5248, pp. 266–282. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
29. Knauff, M., Rauh, R., Renz, J.: A cognitive assessment of topological spatial relations: Results from an empirical investigation. vol. 1329, pp. 193–206. Springer (1997)
30. Kordjamshidi, P., Hois, J., van Otterlo, M., Moens, M.F.: Learning to interpret spatial natural language in terms of qualitative spatial relations. In: Representing Space in Cognition. Explorations in Language and Space, Oxford University Press (2013)

31. Mark, D.M., Comas, D., Egenhofer, M.J., Freundschuh, S.M., Gould, M.D., Nunes, J.: Evaluating and refining computational models of spatial relations through cross-linguistic human-subjects testing. In: Frank, A.U., Kuhn, W. (eds.) Spatial Information Theory A Theoretical Basis for GIS, Lecture Notes in Computer Science, vol. 988, pp. 553–568. Springer, Berlin, Heidelberg (1995)
32. Moratz, R., Tenbrink, T.: Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. Spatial cognition and computation **6**(1), 63–107 (2006)
33. Pradhan, S.S., Ward, W., Martin, J.H.: Towards robust semantic role labeling. Computational linguistics **34**(2), 289–310 (2008)
34. Randell, D., Witkowski, M., Shanahan, M.: From Images to Bodies: Modelling and Exploiting Spatial Occlusion and Motion Parallax. In: Proceedings of the 17th international joint conference on artificial intelligence. pp. 57–63 (2001)
35. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. KR **92**, 165–176 (1992)
36. Regier, T., Carlson, L.A.: Grounding spatial language in perception: an empirical and computational investigation. Journal of experimental psychology: General **130**(2) (2001)
37. Renz, J., Rauh, R., Knauff, M.: Towards cognitive adequacy of topological spatial relations. In: Spatial Cognition II, pp. 184–197. Springer (2000)
38. Richard-Bollans, A., Gómez Álvarez, L., Cohn, A.G.: The Role of Pragmatics in Solving the Winograd Schema Challenge. In: Proceedings of 13th International Symposium on Commonsense Reasoning. CEUR Workshop Proceedings (2017)
39. Rips, L.J.: The Current Status of Research on Concept Combination. Mind & Language **10**(1-2), 72–104 (1995)
40. Rodrigues, E., Santos, P.E., Lopes, M.: Pinning down polysemy: A formalisation for a Brazilian Portuguese preposition. Cognitive Systems Research **41**, 84–92 (2017)
41. Rohde, H., Seyfarth, S., Clark, B., Jäger, G., Kaufmann, S.: Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In: Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue. pp. 107–116 (2012)
42. Schwering, A.: Evaluation of a semantic similarity measure for natural language spatial relations. In: Proceedings of the International Conference on Spatial Information Theory. pp. 116–132. Springer (2007)
43. Speer, R., Havasi, C.: Representing General Relational Knowledge in ConceptNet 5. In: LREC. pp. 3679–3686 (2012), http://ai2-s2-pdfs.s3.amazonaws.com/1b97/b4623cf2f183340e548e0aa53abf0f2963d8.pdf
44. Strube, G.: The role of cognitive science in knowledge engineering. In: Schmalhofer, F., Strube, G., Wetter, T. (eds.) Contemporary Knowledge Engineering and Cognition. pp. 159–174. Springer Berlin Heidelberg, Berlin, Heidelberg (1992)
45. Suchan, J., Bhatt, M.: Commonsense Scene Semantics for Cognitive Robotics. In: ViPAR (2017)
46. Zadeh, L.A.: Fuzzy sets. Information and control **8**(3), 338–353 (1965)
47. Zlatev, J.: Spatial semantics. The Oxford handbook of cognitive linguistics pp. 318–350 (2007)
48. Zwarts, J.: Spatial semantics: Modeling the meaning of prepositions. Language and Linguistics Compass **11**(5) (2017)
49. Zwarts, J., Gärdenfors, P.: Locative and Directional Prepositions in Conceptual Spaces: The Role of Polar Convexity. Journal of Logic, Language and Information **25**(1), 109–138 (2016)

Social Choice and the Problem of Recommending Essential Readings

Silvan Hungerbühler Haukur Páll Jóhnsson Grzegorz Lisowski and Max Rapp

ILLC, Universiteit van Amsterdam

Abstract. We tackle the practical problem of finding a good rule to recommend a collective set of *news items* to a group of media consumers with possibly very disparate individual interest in the available items. For our analysis, we adapt a formal framework from voting theory in *Computational Social Choice* to the media setting in order to compare the performance of five recommendation rules with respect to several desirable properties of recommendation sets. Through simulations, we find that polarization of the audience limits how well these rules can perform in general. On the other hand, greater diversity or universality can be achieved at only low cost in utility.

1 Introduction

How to balance the media's core function of providing news that is relevant to society at large against the increasing economic necessity of offering an individually tailored product? News media face a dilemma: Either submit to highly personalized news feeds on online social media networks that drive political fragmentation, partisanship and contribute to the erosion of society's commonly accepted factual base; or risk losing disgruntled readers, who feel that the issues which they consider important are inadequately represented in the mainstream media, to less reliable Internet news outlets.

Common recommender systems such as matrix factorization algorithms create highly individualized rankings over items based on the users' past behavior (Jannach, Resnick, Tuzhilin, & Zanker, 2016). Could those rankings be aggregated in a principled way to generate a common set of *essential readings* for the whole user group?

The present paper takes a step towards addressing this problem by designing and testing a number of such aggregation rules for news articles using the tools of Social Choice Theory. All that is needed for the rule to work is an ordering of the news articles from first to last according to their importance for each agent. The way in which this preference ordering is elicited from the individual is left open; depending on the concrete application, the data can be thought of as output of a recommender system as suggested above, but could also be explicitly provided by the consumers or gathered by data mining techniques.

Naturally, there are certain properties one would expect such a collection of essential articles to have. The total length of recommended articles for a

newspaper's title page, for example, should not exceed its character limit which relates to a problem of making collective choices with a restricted budget (Lu & Boutilier, 2011). Likewise, there are relations between the essential articles and the rankings by the individuals one would like to see respected by a recommendation rule. For instance, if all consumers detest a certain news item, then it should certainly not be featured in the essential collection instead of another item prioritized by everybody.

This paper aims at better understanding of collective recommendation rules in media settings by formally studying the interaction between rules and properties of their recommendations. We suggest performance metrics to analyze benefits and drawbacks of various ways to determine a set of essential news items for a group, given each member's individual preferences over said items. We proceed by running simulations to estimate the performance of the aggregation rules according to those metrics.

The paper is structured as follows: in Section 2 we provide the formal definition of the recommendation problem as we want to study it. In Section 3 we propose and formally present desirable properties a collection of recommended articles ought to have. In Section 4 we propose five rules for the task of turning individual preferences into a single recommendation. Section 5 contains the methodology, presentation and discussion of our simulation results and suggests directions for future work while Section 6 concludes.

2 Formal Framework

This section specifies the formal framework we use. There is a set of *news items* $A = \{a_1, \ldots, a_m\}$, a subset of which are the *recommended items* $W \subseteq A$ for a group of *consumers* $N = \{1, \ldots, n\}$. Each consumer $i \in N$ has preferences over A represented by a strict, total order $\succ_i$. Let $\mathcal{L}(A)$ be the set of such orders. Then the preference orders of a set of consumers N over news items A form a *profile of preferences* $\mathcal{R} \in \mathcal{L}^n$.

Depending on the context, the cost could be interpreted as the time it takes to read an article, the cognitive resources it takes a consumer to digest it or simply character length. Consequently, each item in A is assigned a specific *cost* by a function $C : A \to \mathbb{N}$. Notice that the notion of *cost* can be straightforwardly generalized to sets of articles. Namely, the cost of a recommendation set $W \subseteq A$ is given by:

$$C(W) = \sum_{w \in W} C(w)$$

In addition, $C^m(A)$ denotes the Cartesian product over $C(a_1) \times ... \times C(a_m)$. Furthermore, we consider the *utility* that the inclusion of an article in the recommendation set gives to particular readers. We follow Lu and Boutilier (2011) in deriving pseudo-utilities from readers' preference orders. For present purposes we used the Borda score, that is, u_i outputs the value $m-1$ for consumer i's top item, $m-2$ for the second one and so forth. This choice allows for simplicity of

the considered setting. It provides a straightforward conversion of rankings over objects to their utility for users. This approach can be useful when a designer of a recommendation system only has information about rankings of options for particular users, not about the extent to which agents would desire them. Other ways of defining utilities are also compatible with our framework. Their exploration would be an interesting direction for future research.

The *utility for a consumer* i is given by a function:

$$u_i : \mathcal{L}^n \times N \times A \to \mathbb{N}$$

Similarly to the cost of articles, their utility can also be generalized to sets of items. The *total utility* of a recommendation set W amounts to:

$$u(W) = \sum_{a \in W} u(a)$$

where $u(a) = \sum_{i=1}^{n} u_i(a)$. In addition, $C^m(A)$ denotes the Cartesian product over $C(a_1) \times ... \times C(a_m)$.

Finally, as these resources are limited we assume a *budget* $B \in \mathbb{N}_{\geq 0}$. In addition, for any $B \in \mathbb{N}_{\geq 0}$, denote by $\mathcal{W}_B$ the set of all elements of $\mathcal{P}(A)$ s.t. $C(W) \leq B$.

Given the notions provided above, we can formulate the definition of a *recommendation rule*. The recommendation rule then is a function from profiles, cost and the budget to recommended items:

$$F : \mathcal{L}^n \times \mathbb{N}^m \times \mathbb{N}_{\geq 0} \to 2^A$$

3 Performance Metrics

We study what properties might justify calling a recommendation "essential". The desired properties of functions selecting a number of options from a set, such as those provided by Elkind, Faliszewski, Skowron, and Slinko (2017), are often binary. A specific function either satisfies them, or it does not. Such properties are standardly referred to as *axioms*. Compared to the sinlge winner case, axioms on multiwinner rules tend to be less salient and enlightening. Therefore, in this work we chose to study properties which functions might satisfy to a certain degree.[1] This choice allows for a more robust comparison between performance of different functions. We refer to those properties as *metrics*.

3.1 Utility Maximization

This metric follows the long tradition arguing in favor of utilitarian social welfare functions (e.g. (Mill, 1874)). Recall that $\mathcal{W}_B$ is the set of all elements of $\mathcal{P}(A)$ s.t. $C(W) \leq B$. A recommendation set W satisfies Utility Maximization iff:

[1] We also obtained results of the axiomatic variant but chose to omit them to be able to give due space to the results presented here.

$$W \in \arg\max_{W' \in \mathcal{W}_B}(u(W'))$$

The motivation behind this property is that, arguably, a recommendation set should get consumers the highest possible payoff. For example, highly popular items about viral memes or the latest celebrity scandals would be favored under a rule that maximizes utility. Even if it does not, it is of interest to assess how far away it is from the optimum. When investigating rules along various other performance dimensions, this allows us to determine the price in terms of utility of improving recommendation sets with respect to those metrics.

3.2 Gini-Coefficient

The Gini-coefficient is the most commonly used measure of inequality in a population. There are many equivalent definitions of the Gini-coefficient (Yitzhaki & Schechtman, 2012, Chapter 2), we define it here in terms of the mean absolute difference between utilities, i.e.:

$$G(W) = \frac{\mathbb{E}[|u_i - u_j|]}{2|W|^{-1}u(W)} = \frac{\frac{1}{|W|^2}\sum_{i=1}^{|W|}\sum_{j=1}^{|W|}|u_i - u_j|}{2|W|^{-1}u(W)} = \frac{\sum_{i=1}^{|W|}\sum_{j=1}^{|W|}|u_i - u_j|}{2|W|u(W)}$$

The Gini-coefficient ranges from 0 (perfect equality, i.e. everybody has the same amount of utility) to 1 (perfect inequality, one agent has all the utility). We care about the Gini-coefficient since unequal distributions of utility increase the likelihood that the worst-off consumers lose interest. But desertion of too large a part of the audience would defeat the point of a common recommendation set. Instead the goal might be to keep everyone just happy enough to keep engaged.

For example, construing the recommendation set as the set of characters on a popular TV show, a producer might face the choice, which, if any, of the characters they should kill off. Killing nobody may take away from the show's suspense. Killing too popular a character may outrage the character's fans too the point of losing interest in the show. Instead, hurting everybody a little may just be the best option. Recommendation sets with a low Gini-coefficient are thus arguably preferable to highly unequal ones.

3.3 Condorcet-Based Metrics

The next two metrics are inspired by the influential Condorcet-criterion for assessing voting rules. The intuition behind the Condorcet method is that pairwise contests should determine winners in an election. An item a is a *Condorcet-winnner* iff it beats every other item b in a majority contest. I.e., a is ranked above b on a majority of the preference orders in the profile for every $b \in A$. There are always either one or zero Condorcet-winners. A voting rule satisfies

the *Condorcet-criterion* iff it always elects the Condorcet-winner if it exists. *Condorcet-extensions* are rules that satisfy the Condorcet-criterion and elect some other item if the Condorcet-winner does not exist. In a multiwnner setting like ours the question is how to select further winners once the Condorcet winner has been added to the recommendation set. The two metrics presented here are two ways to do this in a way that is generalizes the intuition that pairwise contests should determine outcomes.

General Threshold The first metric generalizes the Condorcet method by not only considering the items that win a majority of the *votes* in each contest but also the ones that win a *qualified minority* of votes. Let $N_{a \succ b}$ denote the set of all consumers who rank a over b. We call $\theta \in [0, 1]$ a *general threshold* for a recommendation set W iff

$$a \in W \text{ whenever } \frac{|N_{a \succ b}|}{|N|} \geq \theta \text{ for all } b \in A \setminus \{a\}$$

A recommendation set W is θ-consistent if θ is a general threshold for W. The intuition here is that an item should be in the recommendation set if a qualified minority of the consumers likes the item a lot. The lower the general threshold, the smaller the coalition of consumers needed to predictably push items "on the agenda" as long as they rank those items high enough. If one assumes that an issue's rank corresponds in some sense to how much it affects an individual, the general threshold is a measure of how much importance a recommendation rule assigns to pressing minority issues compared to less salient mainstream topics. E.g., given a profile containing a small group of people at high risk of strokes and a large majority of people suffering from mild headaches, a rule that attempts to optimize the general threshold would favor reports on the stroke issue. Note that the general threshold really is Condorcet-consistent: for any profile, if the Condorcet winner c exists, it will be in W and at $\theta = 0.5$, we have that $W = \{c\}$.

Majority Support Next we consider another generalisation of the Condorcet method. Here we care not only about items that win a majority of *pairwise contests* but also about items that win at least a qualified minority of contests.

Note that the number of pairwise majority contests for a given news item is $|N|-1$. Then $\sigma \in [0, 1]$ is called a majority support (majority support) threshold for a recommendation set W iff

$$a \in W \text{ whenever } \frac{|\{b \in A \setminus \{a\} : \frac{|N_{a \succ b}|}{|N|} > \frac{1}{2}\}|}{|N| - 1} \geq \sigma$$

A recommendation set is σ-consistent if σ is a majority support threshold for W.

In contrast to the general threshold, a low majority support threshold allows consumers to put items on the agenda by establishing a *majority coalition*, even if they do not consider these items as essential. So a rule that minimises the

majority threshold may favor reports on mild headaches over reports on strokes in the previous example.

Again, the Condorcet winner, if it exists, will be in the recommendation set. Furthermore, it should be noted that in the single winner case the majority support rule is known as the Copeland method. In fact, for any σ, any item with a Copeland score (share of majority contests it wins) above σ will also be in W.

Note that utility maximization and Gini-coefficient as well as general threshold and majority support are somewhat complementary: intuitively, a rule which performs well for one of them will have to trade off on the others. When designing a rule one thus has to decide upon a point in the space delimited by the poles optimality and equality on one axis and diversity and universality along another.

This is not an easy task: leaning towards optimality may incur high inequality and desertion of a part of the audience. On the other hand focussing on equality may make for a recommendation set that does not appeal to anyone. Prioritizing universality could lead to neglecting pressing issues that affect only minority groups. Favoring diversity on the other hand risks alienating the average consumer by pushing issues on them which do not affect or interest them.

4 Recommendation Rules

In this section, we study three rules (LGBT, Budgeted Utility Maximization, Budgeted Copeland) that are designed to perform optimally on the metrics utility maximization, general threshold and majority support, respectively, and compare them with two more traditional rules (Budgeted Borda, Budgeted Plurality). Budgeted Utility Maximization is equivalent to the well-known Knapsack problem. Budgeted Colpeland, Borda and Plurality are budgeted versions of the well-known k-multiwinner voting rules (Elkind et al., 2017). The LGBT-rule is novel to the best of our knowledge.

4.1 Extending Multiwinner Rules

We chose to adapt three k-multiwinner voting rules proposed by Elkind et al. (2017) for our setting: A budgeted Plurality rule (Budgeted Plurality) as a baseline, a budgeted Borda rule (Budgeted Borda) as a more sophisticated representative of the positional scoring rules and a budgeted Copeland rule (Budgeted Copeland) to represent the Condorcet extensions. All of these rules assign a score to each item based on the current profile: $S_F : \mathcal{L}^n \times A \to \mathbb{R}$; for brevity we will henceforth refer to all rules that assign a score by *fit-by-score rules* and denote an item's score, given a profile and a rule, as $S_F(a)$, instead of $S_F(\mathcal{R}, a)$.

The *fit-by-score rules* use their respective scores to recommend items in the following way: Start with the complete budget B and put the highest scoring budget-fitting items in the recommendation set. Then do the same for the remaining budget. Continue until the budget is filled. More formally, we define recursively:

$$W_1 = \{\arg\max_{a \in A} S(a)\} \cap \mathcal{W}_B$$

$$W_k = W_{k-1} \cup (\{\arg\max_{a \in A \setminus W_{k-1}} S(a)\} \cap \mathcal{W}_{B-C(W_{k-1})})$$

Then the recommended set is:

$$F(\mathcal{R}, C^m(A), B) = W_{|A|}$$

4.2 Rules Designed for Optimal Performance

It is easy to see that *fit-by-score* combined with the Copeland-Score performs optimally with respect to majority support. It elects the winning set with the lowest possible majority support threshold σ. Intuitively, this means that it enables majority coalitions to kick an item off the agenda even if they dislike that item only a bit. Similarly, we designed rules to perform optimally with respect to utility maximization and general threshold.

Lowest General Budget-compatible Threshold (LGBT) The first such rule is the *Lowest General Budget-compatible Threshold Rule* (LGBT). It is designed to yield optimal results with respect to general threshold. To achieve this, we start by defining an item's θ-score as follows:

$$\theta(a) = \min_{b \in A \setminus \{a\}} \frac{|N_{a \succ b}|}{|N|}$$

Then LGBT applies the *fit-by-score* method to elect the recommended set.

LGBT recommends the set that is optimal for minority preferences in the sense that it chooses the recommendation set with the lowest possible general threshold, thus enabling comparatively small coalitions to push the issues they consider important onto the agenda. For this it is sufficient but not necessary that a plurality of $\theta|N|$ voters rank the desired item first.

Budgeted Utility Maximization The second novel recommendation rule is the *Budgeted Utility Maximization*. Budgeted Utility Maximization is designed to perform optimally with respect to utility maximization. It selects the set of articles which has the greatest overall utility for the consumers while fitting the budget.

$$F(\mathcal{R}, C^m(A), B) = \arg\max_{W \in \mathcal{W}_B} u(W)$$

Although computing the Budgeted Utility Maximization recommendation is an NP-hard problem because it corresponds to the *0-1 Knapsack Problem*, there exist algorithms to solve it in pseudo-polynomial time (Kellerer, Pferschy, & Pisinger, 2004). We used such an algorithm for our simulations.

5 Simulations

We investigated how the rules behave with respect to the metrics of performance defined earlier.

5.1 Method

The distribution of individual preference orders in a profile possibly affects the performance of an aggregation rule. Some rule could function well for very homogeneous preferences, while performing poorly when preferences are very fragmented or even polarly oppose each other. Since only a very limited number of naturally occurring preference profile datasets are available, we generated profiles ourselves in order to capture some plausible types of distributions of preferences amongst a group of individuals. Thereafter we automatically checked the performance of the considered rules with respect to the four desirable properties: Utility Maximization, Gini-Coefficient, General Threshold and Majority Support. The performance was compared with respect to specific profiles.

In order to generate preference profiles as they might naturally occur in different types of consumer populations, we specified seven different *base profiles*. Each of them represented a possible distribution of individual preferences over 10 and 20 items respectively. Every profile contained preference orders of 5000 individual consumers.

The seven base preference orders fall in four categories: *random* profiles — each individual's preference order is sampled independently from the other 4999 preferences in the profile; *fraction* profiles — there are five different clusters of individual preferences which are more similar to each their clustered peers than to the other clusters (there are, in turn, two types of cluster-size distributions: one where the biggest cluster comprises 50% of individual orders and five smaller ones 10% each, and another one where the division is 80% and four times 5%); *polarized* profiles — where two consumer populations have polarly opposed preference orders; and *similar* profiles — where two consumer populations have different preferences orders, but not in a polarized way.

A specific profile, then, is a noisy copy of a base profile. After specifying the base profiles we applied noise to model variety amongst individual consumers. The noise was introduced by employing a probabilistic model to swap items in the preference orders, where farther swaps occurred with smaller probability than closer swaps. Namely, a rank r in the preference order and a swap direction (upwards or downwards) are sampled random-uniformly. Then for each rank r' lower (or higher respectively) than the sampled rank, a swap of the corresponding items is attempted. The success probability of the swap is:

$$\frac{1}{(|r - r'| + 1)^2}$$

This way of introducing noise allows to control the swapping distance between profiles within a cluster by varying the number of ranks to be sampled. It captures the intuition that within a cluster, profiles should be similar to each

other. In this fashion we generated 100 profiles each for the cases of 10 and 20 news items for all seven profile-types.

In terms of the results presented below, the most important base profiles are the following two with two clusters of underlying preferences each. They are meant to account for the differences in the performance of rules in situations when consumers' opinions are homogeneous, and when they are diversified.

For the first, the two clusters are simply noisy copies of a third underlying preference order. We randomly generated one preference order, applied high noise to it twice to get two different preference orders and then generated 2500 moderately noisy copies of both. This resulted in profiles with two *clusters* of fairly similar consumer groups whose orderings agree mostly, but not completely, amongst themselves. We call them *similar-cluster profiles*. For the other base profile with two underlying preference orders, we randomly generated a preference order and created 2500 noisy copies of it. We then reversed the initial ordering and generated 2500 noisy copies of it as well. We call the result *polarized-cluster profiles*.

5.2 Results

We present here our most salient results which pertain to a) how different profiles affect performance on average with respect to our desired properties *for all of our rules* and b) how the rules perform *against each other* on the generated profiles. This allowed us to test for differences across distinct profiles for each rule.

To test the significance of the obtained results, we employed ANOVA in combination with Tukey post hoc analysis.

Results of Type a) Firstly, concerning utility maximization, we found that for all profiles with 20 candidates and two clusters of consumers, the rules performed on average 5 to 10 percentage points worse on polarized-clusters populations than on similar-clusters populations.

Secondly, with respect to general threshold and profiles with two clusters of consumers, all rules perform between 30 and 55 percentage points worse on the polarized-cluster populations than on the similar-cluster populations. This can be seen in Figure 1. For LGBT, these differences were significant on a 95% confidence level for both the 10 and 20 items case. For Budgeted Borda, Budgeted Copeland and Budgeted Plurality they were only significant for the profiles with 10 items.

Results of Type b) We found that both Budgeted Borda and Budgeted Plurality perform significantly worse with respect to utility maximization than Budgeted Utility Maximization on the 95% confidence level. While Budgeted Copeland and LGBT performed worse on utility maximization on average, the difference to Budgeted Utility Maximization was not significant.

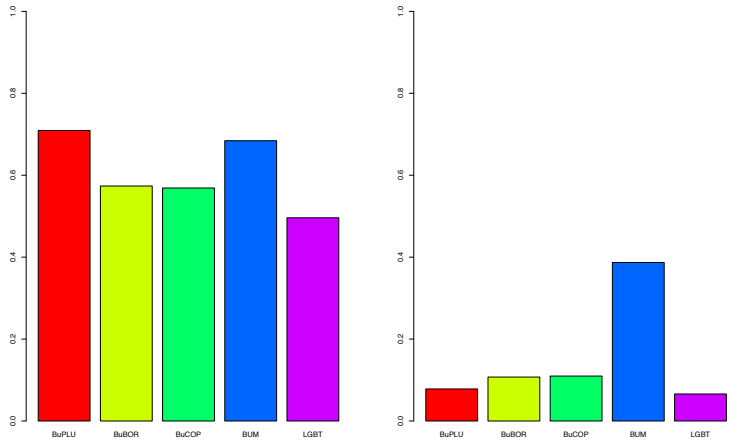


Fig. 1. From left to right the bars stand for Budgeted Plurality, Budgeted Borda, Budgeted Copeland, Budgeted Utility Maximization and LGBT. The left plot shows the average general threshold for polarized profiles. The right plot shows average general threshold for similar profiles.

With respect to general threshold, we found that for both for the populations with 10 and 20 items, the LGBT rule performs significantly better than the other rules on the 99%-level. The difference is most pronounced for the Budgeted Utility Maximization rule. For 10 items Budgeted Utility Maximization scores on average 19.5 percentage points higher (worse) than LGBT. In the case of 20 items, Budgeted Utility Maximization scores on average 21.8 percentage points higher (worse) than LGBT.

For the Gini-coefficient we found that for 20 and 10 items, respectively, Budgeted Utility Maximization performs 3.8 and 10.6 percentage points worse than LGBT, with significance on the 99% level. There were no significant differences between LGBT and the other score-and-fit rules.

Another result pertains to majority support. As expected, Budgeted Copeland performed best with respect to this metric and significantly so compared to all other rules on the 99% interval. Less obviously, we also found that Budgeted Borda outperforms the remaining rules with a significance on the 99% level. Thus, both Budgeted Utility Maximization and LGBT only did approximately as well as the baseline Budgeted Plurality rule with respect to majority support.

5.3 Discussion

The a)-type results indicate that polarization of the readership leads to both lower utility and higher general threshold indicating what one might call a *cost*

of polarization. The cost of polarization is bad news for hopes to find common recommendations for divided audiences. An avenue for future research could be the search for a recommendation rule that minimizes the cost of polarization.

Concerning type b) results, we conclude that no single rule performs optimally with respect to all desirable properties and types of profiles. As expected we have a "three-way-tie" between Budgeted Utility Maximization, LGBT and Budgeted Copeland when it comes to utility maximization, general threshold and majority support. However, looking at the other properties, a tendency can be established: While the Budgeted Utility Maximization rule performs optimally on utility maximization, the difference to Budgeted Copeland and LGBT was not significant which indicates that Budgeted Copeland and LGBT come close to the optimum. A further point to make is that while LGBT and Budgeted Copeland are tied with respect to Gini-coefficient, they clearly beat Budgeted Utility Maximization in this respect.

To sum up: although Budgeted Utility Maximization maximizes utility by construction, this only leads to little added utility compared to Budgeted Copeland and LGBT. Moreover, this additional utility comes at the cost of markedly unequal utility distribution amongst the consumers, higher lowest general threshold and higher lowest majority support thresholds. All of this arguably indicates that Budgeted Utility Maximization is overall a worse rule than Budgeted Copeland and LGBT: higher diversity or greater universality can be achieved at a fairly low price in overall utility. This indicates that greater individualization is not the only way forward when attempting to capture and maintain an audience. Thus maybe the media are not doomed to ever greater splintering and catering to special interests if essential readings are chosen prudently.

In the present paper, we have suggested and tested several rules to carry out this choice. However, in doing so we made some strong assumptions, most notably the equi-distance between neighboring items' utilities. In addition, our results are based purely on simulations. Hence it would be a natural next step to investigate whether our simulation results carry over to an experimental setting with utility data submitted by real media consumers, such as the users of an online news-aggregator. In this setting, one could then also track how a common set of essential readings affects users' preference orders over time. This would also allow to research strategic behavior in the considered framework. Then, we might identify recommendation rules which are the hardest to manipulate.

6 Conclusion

In this paper, we presented a formal Social Choice framework to recommend a common winner set to a group of consumers given a budget constraint. Interpreting the items as news articles, the voters as readers and the budget as e.g. readers' attention span or space on a frontpage this task can be understood as finding a principled way to balance individualization and newsworthiness by selecting a set of *essential readings* to be recommended to all readers. We devised novel performance metrics suitable to this setting which provide different ways

to evaluate the recommended set. With these criteria in mind, we introduced five rules, one of them novel, designed to perform especially well. For all of the rules, we ran simulations in order to assess the rules against our performance measures. For the simulations we defined multiple population-types and for each type, we generated multiple profiles. We then applied our voting rules to the profiles and performed statistical analysis on the results.

Our conclusion is that using Social Choice theory offers an interesting avenue to improve upon existing recommender systems. Depending on one's value judgements, Budgeted Copeland, LGBT and Budgeted Utility Maximization perform well at generating principled common recommendations from individual preference orders. Notably however, the Budgeted Utility Maximization rule only achieves marginal improvement in utility at the expense of diversity, universality and equality. Thus we conclude that LGBT and Budgeted Copeland are the best rules in this setting known so far.

On the other hand we found that polarization of the audience limits what can be achieved by any of these rules. The cost of polarization thus presents an open challenge for designing recommendation rules in the presented setting.

References

Elkind, E., Faliszewski, P., Skowron, P., & Slinko, A. (2017). Properties of Multiwinner Voting Rules. *Social Choice and Welfare*, *48*(3), 599–632.

Jannach, D., Resnick, P., Tuzhilin, A., & Zanker, M. (2016). Recommender Systems — Beyond Matrix Completion. *Commun. ACM*, *59*(11), 94–102.

Kellerer, H., Pferschy, U., & Pisinger, D. (2004). *Knapsack problems*. Springer.

Lu, T., & Boutilier, C. (2011). Budgeted Social Choice: From Consensus to Personalized Decision Making. *IJCAI*, *11*, 280–286.

Mill, J. (1874). *Utilitarianism*. Longmans, Green, Reader&Dyer.

Yitzhaki, S., & Schechtman, E. (2012). *The Gini Methodology: A Primer on a Statistical Methodology*. Springer New York.

Rule-based Reasoners in Epistemic Logic

Anthia Solaki

ILLC, University of Amsterdam
a.solaki2@uva.nl

Abstract. In this paper, we offer a balanced response to the problem of logical omniscience, whereby agents are modeled as non-omniscient yet still logically competent reasoners. To achieve this, we account for the deductive steps that form the epistemic state of an agent. In particular, we introduce operators for applications of inference rules and design a possible-worlds model which is (a) equipped with a syntactic valuation, determining the agent's (explicit) knowledge, and (b) suitably structured by rule-induced transitions between worlds. As a result, we obtain a detailed analysis of the agent's reasoning processes. We then offer validities that exemplify how the problem of logical omniscience is avoided and compare our response to others in the literature. A sound and complete axiomatization is also provided. We finally show how simple extensions of this setting make it compatible with tools from Dynamic Epistemic Logic (DEL) and open to the incorporation of empirical findings on human reasoning.

Keywords: Rule-based reasoners. Epistemic logic. Dynamic epistemic logic. Logical omniscience. Bounded reasoning. Resource-bounded agents. Minimal rationality. Human reasoning.

1 Introduction

Standard (S5) epistemic logic, using possible-worlds semantics, suffers from the *problem of logical omniscience* ([13]): agents are modelled as reasoners with unlimited inferential power, always knowing whatever follows logically from what they know. This stark contrast with reality is also witnessed by experimental results indicating that subjects are systematically fallible in reasoning tasks ([21, 22]). It is even from a normative view that the standard account is insufficient, for it disregards the underlying reasoning of the agent and thus the restrictions on what can be *feasibly* asked of her. Therefore, knowledge should not be subject to logical closure principles. This, however, need not entail that agents are logically incompetent. While we often fail in complex inferences (e.g. due to lack of resources), we do engage in bounded reasoning: knowing that it is raining, and that we need a raincoat whenever it is raining, we do take a raincoat before leaving home. The empirical data also contributes to the case for logical competence, and as proposed in [9], we should seek a standard of *Minimal Rationality*. Drawing on these, we aim at modelling how an agent should *come to know* whatever can be feasibly reached from her epistemic state.

In the twofold project of modelling a non-omniscient yet competent agent, we take on board the observations found in [7]. The deductive steps underpinning knowledge should be clearly reflected in an epistemic framework and this should still be compatible with "external" informational acts, as studied in DEL. We also place another desideratum: in principle, we should be able to employ empirical facts provided by cognitive scientists.

While many attempts have dealt with logical omniscience, not every attempt pursues a solution along the lines just described. Rule-based approaches, mainly applied on Artificial Intelligence, have paved the way towards our direction. Konolige ([16]) uses *belief sets* closed under an (incomplete) set of inference rules, but such (weaker) closure properties do not suffice to capture the agent's reasoning nor its cognitive load. Similar remarks apply to attempts which use modalities for reasoning processes ([11]), state-transitions due to inference ([2, 3, 4]), or arbitrary rule applications ([14]). Collapsing reasoning processes to a modality, without a detailed analysis of their composition, would not help us determine what eventually makes them halt nor exploit investigations in psychology of reasoning which usually study *individual* inference rules on the grounds of cognitive difficulty. Interestingly, in [17], the author develops a logic where rules, accompanied by cognitive costs, are explicitly introduced in the language, but he gives no semantics, rendering the effect of his rule-operators unclear and the choice of axioms controversial.[1] Awareness settings ([12]) discern implicit and explicit attitudes, avoiding omniscience with respect to the latter, which additionally ask that agents are *aware* of a formula. Yet, an arbitrary syntactic awareness-filter cannot be associated with logical competence, and even if ad-hoc modifications are imposed (e.g. awareness closure under subformulas), forms of the problem are retained.[2]

The remainder is organized as follows: we first present our basic setting and explain how it contributes to the solution of the problem (Section 2). We then give a sound and complete axiomatization in Section 3 and in Section 4, we discuss how the basic framework can be easily adjusted to accommodate other directions and include sophisticated tools from logic and cognitive science.

2 The setting

We first construct our logical language, building on the following definitions:

Definition 1 (Inference rule). *Given $\phi_1, \ldots, \phi_n, \psi$ in the standard propositional language $\mathcal{L}_P$ (based on a set of atoms Φ), an inference rule R_i is a formula of the form $\{\phi_1, \ldots, \phi_n\} \rightsquigarrow \psi$.*

[1] In [18] an impossible-worlds semantics is presented, but again reasoning is captured via modalities standing for a *number* of steps; this raises concerns analogous to the ones discussed before.

[2] A notable exception where awareness is affected by reasoning is given in [23]; in what follows, we design a rule-based approach but without appealing to a notion of awareness.

We then use $pr(R_i)$ and $con(R_i)$ to abbreviate the set of premises and the conclusion of R_i.[3] The rule is to say that whenever the premises are true, the conclusion is also true. We also use $\mathcal{L}_{\mathcal{R}}$ to denote the set of inference rules and $\mathcal{L} := \mathcal{L}_P \cup \mathcal{L}_{\mathcal{R}}$.

Definition 2 (Translation). *The translation of a formula in $\mathcal{L}$ is defined as: $Tr(\phi) := \phi$, if $\phi \in \mathcal{L}_P$ and $Tr(R_i) := \bigwedge_{\phi \in pr(R_i)} \phi \to con(R_i)$, if $R_i \in \mathcal{L}_{\mathcal{R}}$.*

We now define the language of this framework:

Definition 3 (Language $\mathcal{L}_{\mathrm{RB}}$). *Given a countable set of propositional atoms Φ, the language $\mathcal{L}_{\mathrm{RB}}$ is defined inductively as follows:*

$$\phi \quad ::= \quad p \quad | \quad \neg\phi \quad | \quad \phi \wedge \phi \quad | \quad K\psi \quad | \quad \langle R_i \rangle \phi$$

with $p \in \Phi, \psi \in \mathcal{L}, R_i \in \mathcal{L}_{\mathcal{R}}$.

As usual, $K\psi$ reads "the agent knows ψ". $\mathcal{L}_{\mathrm{RB}}$ includes knowledge assertions for *rules* too. That is, apart of knowledge of facts, we can also express which *rules* the agent knows (and is therefore capable of applying). Each $\langle R_i \rangle$ is seen as a labeled operator for a rule-application. A formula $\langle R_i \rangle \phi$ reads "after some application of inference rule R_i, ϕ is true". Dual modalities of the form $[R_i]$ such that $[R_i]\phi$ expresses "after *any* application of R_i, ϕ is true", and the remaining Boolean connectives are defined as usual.

Next, we define our model motivated by the idea that reasoning steps, expressed through rule-applications, should be hardwired in it. We introduce possible worlds that are connected according to the effect of inference rules. Since an agent's reasoning affects the information she holds (rather than truth of facts), the usual valuation function is accompanied by a function yielding which formulas the agent knows at each world. In this sense, each world represents what is explicitly known at it and each rule triggers suitable transitions between them.

Definition 4 (Model). *A model is a tuple $M = \langle W, T, V_1, V_2 \rangle$ where*

- *W is a non-empty set of worlds.*
- *$T : \mathcal{L}_{\mathcal{R}} \to \mathcal{P}(W \times W)$ is a function such that a binary relation on W is assigned to each inference rule in $\mathcal{L}_{\mathcal{R}}$. That is, for $R_i \in \mathcal{L}_{\mathcal{R}}$, $T(R_i) = T_i \subseteq W \times W$, standing for the transition between worlds induced by the rule R_i.*
- *$V_1 : W \to \mathcal{P}(\Phi)$ is a valuation function assigning a set of propositional atoms to each world; intuitively those that are true at the world.*
- *$V_2 : W \to \mathcal{P}(\mathcal{L})$ is a function assigning a set of formulas of $\mathcal{L}$ to each world; intuitively those that the agent knows at the world.*

The truth clauses are given as follows:

[3] We emphasize that R_i denotes a *single* rule instance. The rule, which is in fact a pair, composed of the set of premises and the conclusion, is given in terms of the notation $\rightsquigarrow$ for readability and convenience.

Definition 5 (Truth clauses).

- $M, w \models p$ *if and only if* $p \in V_1(w)$ *for* $p \in \Phi$.
- $M, w \models K\phi$ *if and only if* $\phi \in V_2(w)$.
- $M, w \models \neg\phi$ *if and only if* $M, w \not\models \phi$.
- $M, w \models \phi \wedge \psi$ *if and only if* $M, w \models \phi$ *and* $M, w \models \psi$.
- $M, w \models \langle R_i \rangle \phi$ *if and only if there exists some* $u \in W$ *such that* wT_iu *and* $M, u \models \phi$.

A formula is *valid in a model* if it is true at every world of the model and *valid* if it is valid in the class of all models. However, certain conditions have to be imposed on our initial, general class, to capture the desired effect of rule-applications. To that end, we need the following:

Definition 6 (Propositional truths). *Let* M *be a model and* $w \in W$ *a world of the model. Its set of propositional truths is* $V_1^*(w) = \{\phi \in \mathcal{L}_P \mid M, w \models \phi\}$.

We can now fix an appropriate class of models, denoted by $\mathbf{M}$. For any model M (with $T(R_i) = T_i$ as defined above), $M \in \mathbf{M}$ if and only if:

1. For any inference rule $R_i = \{\phi_1, \ldots, \phi_n\} \rightsquigarrow \psi$, if $w \in W$ is such that $R_i \in V_2(w)$ and $\phi_1, \ldots, \phi_n \in V_2(w)$, then there exists a world $u \in W$ such that wT_iu.
2. For any $w, u \in W$ and inference rule $R_i = \{\phi_1, \ldots, \phi_n\} \rightsquigarrow \psi$, if wT_iu then $R_i \in V_2(w)$, $\phi_1, \ldots, \phi_n \in V_2(w)$ and $V_2(u) = V_2(w) \cup \{\psi\}$.
3. For any $w \in W$ and $\phi \in \mathcal{L}$, if $\phi \in V_2(w)$ then $Tr(\phi) \in V_1^*(w)$.
4. For any $w, u \in W$ and inference rule R_i, if wT_iu then $V_1^*(w) = V_1^*(u)$.

Condition 1 says that if a world represents an epistemic state containing the premises of a known rule R_i, then it must be connected to some other world by the corresponding T_i. Condition 2 says that if w is T_i-connected to u, then it must be that u enriches the epistemic state of w in terms of R_i. This is to ensure that each transition is associated with some addition of a conclusion to an epistemic state. Condition 3 is imposed to guarantee the veridicality of knowledge and the soundness of the known rules.[4] Finally, condition 4 states that T_i-connected worlds are propositionally indiscernible, i.e. transitions stand for purely epistemic actions.

We present some validities that illustrate desirable properties of reasoning processes and will be instrumental for a balanced response against logical omniscience. For notational convenience, we abbreviate sequences of rules as follows:

- $\langle \ddagger \rangle := \langle R_1 \rangle \ldots \langle R_n \rangle$
- $\langle \dagger \rangle := \langle R'_1 \rangle \ldots \langle R'_m \rangle$

[4] Recall that $V_2 : W \to \mathcal{P}(\mathcal{L})$ and that $\mathcal{L} := \mathcal{L}_P \cup \mathcal{L}_{\mathcal{R}}$. Moreover, it should be clear that the world u whose existence is guaranteed by condition 1, is such that it contains the conclusion of R_i, by condition 2, and the rule R_i is necessarily sound due to condition 3.

standing for "after some application of $R_1(R'_1)$, followed by some application of $R_2(R'_2), \ldots$, followed by some application of $R_n(R'_m)$" (in that order). Similar abbreviations can be defined for the dual cases; for example, by using $[R_1], \ldots, [R_n]$ for the first sequence and $[R'_1], \ldots, [R'_m]$ for the second.

Theorem 1 (M-validities).

1. $\langle\ddagger\rangle K\phi \to Tr(\phi)$ *is valid in the class* $\boldsymbol{M}$. *(Factivity)*
2. $\langle\ddagger\rangle K\phi \to \langle\ddagger\rangle[\dagger]K\phi$ *is valid in the class* $\boldsymbol{M}$. *(Persistence)*
3. $\langle\ddagger\rangle K\phi \wedge \langle\dagger\rangle K\psi \to \langle\ddagger\rangle\langle\dagger\rangle(K\phi \wedge K\psi)$ *is valid in the class* $\boldsymbol{M}$. *(Merge)*
4. *For any inference rule* R_i, $KR_i \wedge \bigwedge_{\phi \in pr(R_i)} K\phi \to \langle R_i\rangle Kcon(R_i)$ *is valid in the class* $\boldsymbol{M}$. *(Success)*

Proof.

1. Take arbitrary model $M \in \mathbf{M}$ and arbitrary world $w \in W$ of the model. Suppose $M, w \models \langle\ddagger\rangle K\phi$. Unpacking the sequence according to the abbreviation, $M, w \models \langle R_1\rangle \ldots \langle R_n\rangle K\phi$, for the inference rules $R_1, \ldots, R_n$. Following Definition 5, there is a world $u_1 \in W$ such that wT_1u_1 and $M, u_1 \models \langle R_2\rangle \ldots \langle R_n\rangle K\phi$. Continuing like that, there is a world $u_n \in W$ such that $u_{n-1}T_nu_n$ and $M, u_n \models K\phi$, which in turn amounts to $\phi \in V_2(u_n)$. Then, by condition 3, $Tr(\phi) \in V_1^*(u_n)$. From condition 4, $Tr(\phi) \in V_1^*(u_{n-1})$. Continuing this process backwards, $Tr(\phi) \in V_1^*(w)$. Therefore $M, w \models Tr(\phi)$. Given the arbitrariness of $M \in \mathbf{M}$ and $w \in W$, we finally conclude that the formula is valid in the class $\mathbf{M}$.
2. Take arbitrary model $M \in \mathbf{M}$ and arbitrary world $w \in W$ of the model. Suppose $M, w \models \langle\ddagger\rangle K\phi$. Unpacking the sequence according to the abbreviation, this amounts to $M, w \models \langle R_1\rangle \ldots \langle R_n\rangle K\phi$. As in the previous case, we obtain a chain $wT_1u_1 \ldots u_{n-1}T_nu_n$ such that $M, u_n \models K\phi$, which in turn amounts to $\phi \in V_2(u_n)$ (1). It suffices to show that $M, u_n \models [\dagger]K\phi$, i.e., by repeating the unpacking, now for $[\dagger] = [R'_1] \ldots [R'_m]$, that for every world $v_1 \in W$ such that $u_nT'_1v_1, \ldots$, for every world $v_m \in W$ such that $v_{m-1}T'_mv_m$, $M, v_m \models K\phi$, i.e. $\phi \in V_2(v_m)$. Take arbitrary such $v_1, \ldots, v_m$. Then due to condition 2 and (1), $\phi \in V_2(v_1)$ and continuing in the same fashion $\phi \in V_2(v_m)$. Therefore, $M, w \models \langle\ddagger\rangle[\dagger]K\phi$, hence $M, w \models \langle\ddagger\rangle K\phi \to \langle\ddagger\rangle[\dagger]K\phi$, as desired.
3. Take arbitrary model $M \in \mathbf{M}$ and arbitrary world $w \in W$ of the model. Suppose $M, w \models \langle\ddagger\rangle K\phi \wedge \langle\dagger\rangle K\psi$. So $M, w \models \langle\ddagger\rangle K\phi$ and $M, w \models \langle\dagger\rangle K\psi$. As above, we obtain a chain $wT_1u_1 \ldots u_{n-1}T_nu_n$ such that $M, u_n \models K\phi$, i.e. $\phi \in V_2(u_n)$, and a chain $wT'_1v_1 \ldots v_{m-1}T'_nv_m$ such that $M, v_m \models K\psi$, i.e. $\psi \in V_2(v_m)$. The rough idea of the proof is to make use of the conditions of $\mathbf{M}$ to merge the two chains. By condition 2, we know that $V_2(w) \subseteq V_2(u_n)$ and that $V_2(w)$ contains all the premises of rule R'_1, as well as the rule itself. Therefore, $V_2(u_n)$ in turn contains all the premises of rule R'_1 and the rule itself. By conditions 1 and 2, there is a world z_1 such that $u_nT'_1z_1$ and $V_2(z_1) = V_2(u_n) \cup \{con(R'_1)\}$. Now again, by condition 2, $V_2(v_1) = V_2(w) \cup \{con(R'_1)\}$ and since $V_2(w) \subseteq V_2(u_n)$: $V_2(v_1) \subseteq V_2(z_1)$, so we know

that z_1 contains the premises for R'_2 and the rule itself. Again by conditions 1 and 2, there is a world z_2 such that $z_1 T'_2 z_2$ and $V_2(z_2) = V_2(z_1) \cup \{con(R'_2)\}$. Continuing like that, the alternations of condition 2 and condition 1, based on the initial assumptions, yield a world z_m such that $z_{m-1} T'_m z_m$ and $V_2(z_m) = V_2(z_{m-1}) \cup \{con(R'_m)\}$ with $V_2(v_m) \subseteq V_2(z_m)$. Therefore $\psi \in V_2(z_m)$. In addition, as the constructed chain is of the form $u_n T'_1 z_1 T'_2 z_2 \ldots T'_m z_m$ and due to condition 2, $\phi \in V_2(z_m)$. So $M, z_m \models K\phi \wedge K\psi$, i.e. $M, u_n \models \langle\dagger\rangle(K\phi \wedge K\psi)$. So finally $M, w \models \langle\ddagger\rangle\langle\dagger\rangle(K\phi \wedge K\psi)$, as desired.

4. Take arbitrary model $M \in \mathbf{M}$ and arbitrary world $w \in W$ of the model. Suppose $M, w \models KR_i \wedge \bigwedge_{\phi \in pr(R_i)} K\phi$. Then $R_i \in V_2(w)$ and $\phi \in V_2(w)$, for every $\phi \in pr(R_i)$. Next, from conditions 1 and 2, there is $v \in W$ such that $w T_i v$ and $V_2(v) = V_2(w) \cup \{con(R_i)\}$. As a result, $M, v \models Kcon(R_i)$. Finally, $M, w \models \langle R_i \rangle Kcon(R_i)$, as desired.

Factivity says that whatever comes to be known is true, i.e. only true information or sound rules become known after reasoning, and *Persistence* says that it remains to be known throughout subsequent reasoning processes. *Merge* exemplifies how the agent merges different reasoning processes, thereby coming to know their outcomes. *Success* captures the effect of applying a rule: the conclusion is added in the agent's epistemic stack. As a concrete example, take the validity of $\bigwedge_{R_i = DNE, MP, CI} KR_i \wedge K\neg\neg\phi \wedge K(\phi \to \psi) \to \langle DNE\rangle\langle MP\rangle\langle CI\rangle K(\phi \wedge \psi)$: after successive applications of specific rules, namely *Double Negation Elimination* ($\{\neg\neg\phi\} \rightsquigarrow \phi$), *Modus Ponens* ($\{\phi, \phi \to \psi\} \rightsquigarrow \psi$) and *Conjunction Introduction* ($\{\phi, \psi\} \rightsquigarrow \phi \wedge \psi$), the agent's knowledge is gradually increased.

Logical omniscience is indeed avoided in a balanced way, i.e. still escaping trivialized, totally ignorant agents. The values of knowledge assertions are determined by V_2, which need not obey any closure principle. On the other hand, suitable applications of inference rules, reflecting the effort to eventually reach a conclusion, ensure that an agent can *come to know* consequences of her knowledge, provided that she follows the appropriate reasoning track. This is how we avoid an implausible commitment to an automatic and effortless way to expand one's knowledge, as the standard validity $K\phi_1 \wedge \ldots \wedge K\phi_n \to K\psi$ would dictate. Besides, Cherniak ([9]) emphasizes that we should view complex deductive reasoning as a task consisting of simple reasoning steps conjoined together. He also argues for a "well-ordering of inferences" in terms of their difficulty, depending both on the rule scheme in question and the logical complexity of its components. Similarly, according to Rips ([20]), deductive reasoning is a psychological procedure in which sets of formulas are connected via links, that essentially amount to applications of inference rules, just as our framework predicts. Overall, competence is preserved because we unfold the actual processes that result in knowledge and account for their dynamic nature. Logical ignorance is thus ruled out because of a more realistic modelling of the underlying reasoning and not because of ad-hoc restrictions imposed on an inflexible notion of knowledge.

It is interesting to see how our rule-based setting fits in the landscape of similar attempts. As in [1, 14], temporal-style connections encode the progress in

the agent's reasoning.[5] Unlike [4, 11, 14, 18], we abstain from a generic notion of reasoning process, instead accounting explicitly for (a) specific rules available to the agent, (b) their individual applications, (c) their chronology, thus monitoring the path that eventually leads to knowledge. This elaborate analysis is, as we remarked above and will further discuss in Section 4, crucial in bridging epistemic frameworks with empirical facts.[6] Furthermore, the enterprise of providing a semantics contributes to Rasmussen's attempt ([17]), who keeps track of rules applied by the agent, on one hand, but lacks a principled way to assess the validity of his proposed axioms, on the other. Constructing a suitable semantic model that reflects rule-based reasoning gives a concrete view on the credibility of axioms and the adequacy of the solution. Finally, implicit and explicit notions can be discerned, not through an arbitrary filter (as with awareness), but through the analysis of the agent's reasoning.

3 Axiomatization

In this section, we develop the logic Λ_{RB}. We thus obtain a full-fledged logical response against the problem and solid ground to defend our selected axioms.

Definition 7 (Axiomatization of Λ_{RB}). *The axiomatization of Λ_{RB} is given by Table 1.*

Table 1.

Axioms	
PC	All instances of classical propositional tautologies
K	$[R_i](\phi \to \psi) \to ([R_i]\phi \to [R_i]\psi)$
T	$K\phi \to Tr(\phi)$
Succession	$KR_i \wedge \bigwedge_{\phi \in pr(R_i)} K\phi \to \langle R_i \rangle \top$
Tracking knowledge	$\langle R_i \rangle K\chi \to \bigwedge_{\phi \in pr(R_i)} K\phi \wedge KR_i \wedge K\chi$, for $\chi \neq con(R_i)$
Knowledge of conclusions	$[R_i]Kcon(R_i)$
$Prop_1$	$\langle R_i \rangle \phi \to \phi$, for $\phi \in \mathcal{L}_P$
$Prop_2$	$\phi \to [R_i]\phi$, for $\phi \in \mathcal{L}_P$
Monotonicity	$K\chi \to [R_i]K\chi$
Rules	
Modus Ponens	From ϕ and $\phi \to \psi$, infer ψ
Necessitation	From ϕ infer $[R_i]\phi$

[5] We note that the frameworks described in [1, 2, 3] that extend the idea of state-transitions to multi-agent settings are particularly interesting for the development of multi-agent variants of our framework too.

[6] More on why this is a worthwhile task can be found in [6].

Theorem 2 (Soundness). *The logic Λ_{RB} is sound with respect to $\boldsymbol{M}$.*

Proof. It suffices to show that the axioms of Definition 7 are valid in the class **M**, as our rules preserve validity as usual.

- The claim for PC and K is trivial.
- The claim for T follows immediately from condition 3.
- The claim for *Succession* follows from condition 1.
- For *Tracking knowledge*: Take any model $M \in \mathbf{M}$ and world $w \in W$ of the model such that $M, w \models \langle R_i \rangle K\chi$, for $R_i = \{\phi_1, \ldots, \phi_n\} \rightsquigarrow \psi$. So there is $u \in W$ such that wT_iu and $\chi \in V_2(u)$. By condition 2, $\phi_1, \ldots, \phi_n, R_i \in V_2(w)$ and since $V_2(u) = V_2(w) \cup \{\psi\}$, $\chi \in V_2(w) \cup \{\psi\}$. So either $\chi \in V_2(w)$ or $\chi = \psi$. Finally, $M, w \models K\phi_1 \wedge \ldots \wedge K\phi_n \wedge KR_i \wedge K\chi$, for $\chi \neq \psi$.
- The claim for *Knowledge of conclusions* follows from condition 2.
- For $Prop_1$: Take any model $M \in \mathbf{M}$ and world $w \in W$ of the model such that $M, w \models \langle R_i \rangle \phi$ for $\phi \in \mathcal{L}_P$. Then, there is $u \in W$ such that wT_iu and $M, u \models \phi$, i.e. $\phi \in V_1^*(u)$. By condition 4, $\phi \in V_1^*(w)$, i.e. $M, w \models \phi$ as desired.
- For $Prop_2$: Take any model $M \in \mathbf{M}$ and world $w \in W$ of the model such that $M, w \models \phi$. Take any $u \in W$ such that wT_iu. Then by condition 4, $\phi \in V_1^*(u)$, i.e. $M, u \models \phi$ so $M, w \models [R_i]\phi$, as desired.
- For *Monotonicity*: Take any model $M \in \mathbf{M}$ and world $w \in W$ of the model such that $M, w \models K\chi$, i.e. $\chi \in V_2(w)$. Take any $u \in W$ such that wT_iu. From condition 2, $\chi \in V_2(u)$, i.e. $M, u \models K\chi$. But then indeed $M, w \models [R_i]K\chi$.

Aiming at completeness, we follow the procedure of [8], employing *canonical models*.

Lemma 1 (Lindenbaum's Lemma). *If Γ is a Λ_{RB}-consistent set of formulas, then it can be extended to a maximal Λ_{RB}-consistent set Γ^+.*

Proof. The proof goes as usual in these cases. After enumerating $\phi_0, \phi_1, \ldots$, the formulas of our language, one constructs the set Γ^+ as $\bigcup_{n \geq 0} \Gamma^n$ where: $\Gamma^0 = \Gamma$, $\Gamma^{n+1} = \Gamma^n \cup \{\phi_n\}$, if this is Λ_{RB}-consistent and $\Gamma^n \cup \{\neg\phi_n\}$ otherwise. The desired properties are easily obtained due to this construction.

Definition 8 (Canonical Model). *The canonical model $\mathcal{M}$ for Λ_{RB} is a tuple $\langle \mathcal{W}, \mathcal{T}, \mathcal{V}_1, \mathcal{V}_2 \rangle$ where:*

- $\mathcal{W} = \{w \mid w$ *a maximal Λ_{RB}-consistent set*$\}$.
- $\mathcal{T} : \mathcal{L}_{\mathcal{R}} \to \mathcal{P}(\mathcal{W} \times \mathcal{W})$, *such that for $R_i \in \mathcal{L}_{\mathcal{R}}$, $\mathcal{T}(R_i) = \mathcal{T}_i$, where $w\mathcal{T}_iu$ if and only if $\{\langle R_i \rangle \phi \mid \phi \in u\} \subseteq w$.*
- $\mathcal{V}_1 : \mathcal{W} \to \mathcal{P}(\Phi)$ *such that* $\mathcal{V}_1(w) = \{p \in \Phi \mid p \in w\}$.
- $\mathcal{V}_2 : \mathcal{W} \to \mathcal{P}(\mathcal{L})$ *such that* $\mathcal{V}_2(w) = \{\phi \in \mathcal{L} \mid K\phi \in w\}$.

It is easy to see that an equivalent formulation for the definition of $\mathcal{T}_i$ is $\{\phi \mid [R_i]\phi \in w\} \subseteq u$. Given the definition of the canonical model and our language $\mathcal{L}_{\mathrm{RB}}$, we show:

Lemma 2 (Existence lemma). *For any formula ϕ in our language and $w \in \mathcal{W}$, if $\langle R_i\rangle\phi \in w$ then there is $u \in \mathcal{W}$ such that $w\mathcal{T}_i u$ and $\phi \in u$.*

Proof. Suppose $\langle R_i\rangle\phi \in w$. Take $S = \{\phi\} \cup \{\psi \mid [R_i]\psi \in w\}$. This set is consistent. Were it inconsistent, there would be $\psi_1, \ldots, \psi_n$ such that $\vdash_{\Lambda_{\mathrm{RB}}} \psi_1 \wedge \ldots \wedge \psi_n \rightarrow \neg\phi$. Using $[R_i]$-necessitation, distribution and propositional tautologies we obtain $\vdash_{\Lambda_{\mathrm{RB}}} ([R_i]\psi_1 \wedge \ldots \wedge [R_i]\psi_n) \rightarrow [R_i]\neg\phi$. By the property of w as maximal consistent set and since $[R_i]\psi_1, \ldots, [R_i]\psi_n \in w$: $[R_i]\neg\phi \in w$. Therefore $\neg\langle R_i\rangle\phi \in w$. Indeed, we have reached a contradiction. Next, we extend S to S^+ according to Lindenbaum's lemma. Then, $\phi \in S^+$ and $[R_i]\psi \in w$ implies $\psi \in S^+$. Take $u := S^+$. As a result, $w\mathcal{T}_i u$ and $\phi \in u$.

Lemma 3 (Truth lemma). *For any formula ϕ in our language and $w \in \mathcal{W}$: $\mathcal{M}, w \models \phi$ if and only if $\phi \in w$.*

Proof. The proof is by induction on the complexity of ϕ.

- Base cases: Consider $\phi := p$ with $p \in \Phi$. Then $\mathcal{M}, w \models p$ if and only if $p \in \mathcal{V}_1(w)$, and by definition, this is the case if and only if $p \in w$. Next, take $\phi := K\psi$ with $\psi \in \mathcal{L}$. Then $\mathcal{M}, w \models K\psi$ if and only if $\psi \in \mathcal{V}_2(w)$, and by definition, this is the case if and only if $K\psi \in w$.
- For $\phi := \neg\psi$ and $\phi := \psi \wedge \chi$, the claim follows easily from I.H. and the maximal consistency of w.
- For $\phi := \langle R_i\rangle\psi$ with I.H. that the result holds for ψ. Then $\mathcal{M}, w \models \langle R_i\rangle\psi$ if and only if there is $u \in \mathcal{W}$ such that $w\mathcal{T}_i u$ and $\mathcal{M}, u \models \psi$. By I.H. this is the case if and only if $\psi \in u$, and by definition of $\mathcal{T}_i$, we get $\langle R_i\rangle\psi \in w$. The other direction follows immediately from the existence lemma.

Theorem 3 (Completeness). *For any set of formulas Γ and formula ϕ in our language: $\Gamma \models_{\boldsymbol{M}} \phi$ only if $\Gamma \vdash_{\Lambda_{\mathrm{RB}}} \phi$.*

Proof.

- We first expand Γ to a maximal Λ_{RB}-consistent set Γ^+. Then, let the canonical model $\mathcal{M}$ as constructed according to Definition 8. Then by Lemma 3, $\mathcal{M}, \Gamma^+ \models \Gamma$. It suffices to show that $\mathcal{M}$ fulfills the conditions of **M**.
- Condition 1 is satisfied.
  Take inference rule $R_i = \{\phi_1, \ldots, \phi_n\} \rightsquigarrow \psi$ and $w \in \mathcal{W}$ with $R_i, \phi_1, \ldots, \phi_n \in \mathcal{V}_2(w)$, i.e. KR_i, $K\phi_1, \ldots, K\phi_n \in w$ (1). We want to show that there is a world $u \in \mathcal{W}$ such that $w\mathcal{T}_i u$. From (1), $KR_i \wedge K\phi_1 \wedge \ldots \wedge K\phi_n \in w$. But from *Succession*, we get that $\langle R_i\rangle\top \in w$. Using the existence lemma, there is indeed $u \in \mathcal{W}$ such that $w\mathcal{T}_i u$.
- Condition 2 is satisfied.
  Suppose that $w\mathcal{T}_i u$ with $R_i = \{\phi_1, \ldots, \phi_n\} \rightsquigarrow \psi$, i.e. if $\phi \in u$ then $\langle R_i\rangle\phi \in w$. Take arbitrary $\chi \in \mathcal{V}_2(u)$. That is, $K\chi \in u$. Therefore, $\langle R_i\rangle K\chi \in w$. From *Tracking knowledge*, $\phi_1, \ldots, \phi_n, R_i \in \mathcal{V}_2(w)$. From *Knowledge of conclusions* and definition of $\mathcal{T}_i$, $K\psi \in u$, i.e. $\psi \in \mathcal{V}_2(u)$. Furthermore by this definition and *Monotonicity* we obtain that $\mathcal{V}_2(w) \subseteq \mathcal{V}_2(u)$. Therefore,

$\mathcal{V}_2(w) \cup \{\psi\} \subseteq \mathcal{V}_2(u)$. Next take $\phi \in \mathcal{V}_2(u)$ with $\phi \neq \psi$. Then $\langle R_i \rangle K\phi \in w$. From *Tracking knowledge*, $K\phi \in w$. As a result, $\phi \in \mathcal{V}_2(w)$. Clearly then, $\mathcal{V}_2(u) = \mathcal{V}_2(w) \cup \{\psi\}$.

– Condition 3 is satisfied.
Let ϕ be a formula in $\mathcal{L}$. Suppose that $\phi \in \mathcal{V}_2(w)$. That is, $K\phi \in w$. Then by T we obtain, $Tr(\phi) \in w$, that is $\mathcal{M}, w \models Tr(\phi)$ and therefore $Tr(\phi) \in \mathcal{V}_1^*(w)$.

– Condition 4 is satisfied.
Take $w, u \in \mathcal{W}$ and $w\mathcal{T}_i u$. By definition of $\mathcal{T}_i$, if $\phi \in u$ then $\langle R_i \rangle \phi \in w$. Now take arbitrary $\phi \in \mathcal{L}_P$ such that $\mathcal{M}, u \models \phi$, i.e. $\phi \in \mathcal{V}_1^*(u)$. This means that $\phi \in u$, therefore $\langle R_i \rangle \phi \in w$. From $Prop_1$, we obtain $\phi \in w$, i.e. $\mathcal{M}, w \models \phi$ so $\phi \in \mathcal{V}_1^*(w)$. As ϕ was arbitrary, $\mathcal{V}_1^*(u) \subseteq \mathcal{V}_1^*(w)$. For the other inclusion, take arbitrary $\phi \in \mathcal{L}_P$ such that $\mathcal{M}, w \models \phi$, i.e. $\phi \in \mathcal{V}_1^*(w)$. This means that $\phi \in w$. From $Prop_2$, we get that $[R_i]\phi \in w$ too. Then we exploit the alternative definition of $\mathcal{T}_i$; since $[R_i]\phi \in w$, $\phi \in u$, i.e. $\mathcal{M}, u \models \phi$ so $\phi \in \mathcal{V}_1^*(u)$. As ϕ was arbitrary, $\mathcal{V}_1^*(w) \subseteq \mathcal{V}_1^*(u)$. Overall, $\mathcal{V}_1^*(w) = \mathcal{V}_1^*(u)$.

4 Extensions

This setting, whose key elements have been hitherto described, can also accommodate more intricate scenarios and facilitate applications informed by other disciplines. In particular, we briefly explain that other tools from (D)EL can be naturally combined with our rule-based logic and that, apart from AI, our syntactic approach can be also relevant for cognitive science.

First, a notion of *implicit* knowledge is not precluded in our framework, for it too employs possible worlds and can be easily endowed with an accessibility relation. Notions of belief can be also included along the lines presented so far, i.e. by simply attaching another function to the model, now yielding the explicit beliefs. Nevertheless, one might drop conditions on factivity or monotonicity. Moreover, just like *public announcements* of DEL,[7] which may enhance the agent's knowledge, there can be actions for the learning of *rules*. Their effect is captured by (suitably) tweaking V_2 so as to add the rule in question. Regarding higher-order knowledge – provided that the language and the range of V_2 are extended – we can also avoid unlimited introspection, as is arguably desired for non-ideal agents. Just as with factual reasoning though, our framework can model *moderate* introspective abilities, via the introduction of introspective rules, whose semantic effect is similarly captured via world transitions.

The use of labeled operators and the order-sensitivity of applications of rules make it easier to exploit the observations of cognitive scientists for a precise modelling of resource-bounded reasoners. For example, [15, 20, 22] suggest that not all rules are equally difficult for agents. According to Rips ([20]), the length and the difficulty of the rules involved in the mental proof constructed for a complex

[7] As usual in DEL ([5, 10]), we can add action operators to our language and capture their effect via model transformations triggered by the action. A formula with dynamic operators, of the form $[\alpha]\phi$, is evaluated by examining what the truth value of ϕ is at a transformed model, obtained via action α.

reasoning task determines its overall difficulty. In [19] the need to assign different weights to different rules is experimentally verified and in [24] empirically calculated weights are attached to different rules. Our framework can take these points into consideration. By fixing the agent's capacity, attaching empirically indicated weights to rules and introducing inequality formulas to the language, we can place preconditions to applications of rules and therefore pinpoint where the cutoff of a reasoning process lies.

On a more technical note, while we have presented a Hilbert-style axiomatization of Λ_{RB}, it would be interesting to develop a labeled sequent calculus alternative to this and investigate the proof-theoretic properties of our system. This investigation can be especially relevant to the state-transition settings studying single- or multi-agent reasoning processes. In this way, we can obtain other independent technical results to motivate the use of such systems.

5 Conclusions

We argued that one of the important challenges for epistemic logic is not only to overcome logical omniscience, but to do so while securing the logical competence of agents. We located this endeavour's key parameter in bounded reasoning and spelled it out in logical terms by keeping track of the inference rules the agent applies. We explained how this enriches existing rule-based approaches and expands the scope of their applications. A sound and complete axiomatization was also provided, followed by a summary of our extensions of the core setting.

References

[1] Ågotnes, T., Alechina, N.: The dynamics of syntactic knowledge. Journal of Logic and Computation 17(1), 83–116 (2007)

[2] Ågotnes, T., Walicki, M.: A logic of reasoning, communication and cooperation with syntactic knowledge. In: AAMAS (2005)

[3] Alechina, N., Jago, M., Logan, B.: Modal logics for communicating rule-based agents. In: ECAI (2006)

[4] Alechina, N., Logan, B.: A logic of situated resource-bounded agents. Journal of Logic, Language and Information 18, 79–95 (2009)

[5] Baltag, A., Renne, B.: Dynamic epistemic logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2016 edn. (2016)

[6] van Benthem, J.: Logic and reasoning: Do the facts matter? Studia Logica: An International Journal for Symbolic Logic 88(1), 67–84 (2008)

[7] van Benthem, J.: Tell it like it is: Information flow in logic. Journal of Peking University (Humanities and Social Science Edition) 1, 80–90 (2008)

[8] Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge University Press, New York, NY, USA (2001)

[9] Cherniak, C.: Minimal Rationality. Bradford book, MIT Press (1986)

[10] van Ditmarsch, H., van der Hoek, W., Kooi, B.: Dynamic Epistemic Logic. Springer Publishing Company, Incorporated (2007)

[11] Duc, H.N.: Reasoning about rational, but not logically omniscient, agents. Journal of Logic and Computation 7(5), 633 (1997)
[12] Fagin, R., Halpern, J.Y.: Belief, awareness, and limited reasoning. Artificial Intelligence 34(1), 39–76 (1987)
[13] Fagin, R., Halpern, J.Y., Moses, Y., Y., V.M.: Reasoning About Knowledge. MIT press (1995)
[14] Jago, M.: Epistemic logic for rule-based agents. Journal of Logic, Language and Information 18(1), 131–158 (2009)
[15] Johnson-Laird, P.N., Byrne, R.M., Schaeken, W.: Propositional reasoning by model. Psychological Review 99(3), 418–439 (1992)
[16] Konolige, K.: A Deduction Model of Belief. Morgan Kaufmann Publishers (1986)
[17] Rasmussen, M.S.: Dynamic epistemic logic and logical omniscience. Logic and Logical Philosophy 24, 377–399 (2015)
[18] Rasmussen, M.S., Bjerring, J.C.: A dynamic solution to the problem of logical omniscience. Journal of Philosophical Logic (forthcoming)
[19] Rijmen, F., De Boeck, P.: Propositional reasoning: The differential contribution of "rules" to the difficulty of complex reasoning problems. Memory & Cognition 29(1), 165–175 (2001)
[20] Rips, L.J.: The Psychology of Proof: Deductive Reasoning in Human Thinking. MIT Press, Cambridge, MA, USA (1994)
[21] Stanovich, K.E., West, R.F.: Individual differences in reasoning: Implications for the rationality debate? Behavioral and Brain Sciences 23(5), 645–665 (2000)
[22] Stenning, K., van Lambalgen, M.: Human Reasoning and Cognitive Science. Boston, USA: MIT Press (2008)
[23] Velázquez-Quesada, F.R.: Small Steps in Dynamics of Information. Ph.D. thesis, Institute for Logic, Language and Computation (ILLC), Amsterdam, The Netherlands (2011)
[24] Zhai, F., Szymanik, J., Titov, I.: Toward probabilistic natural logic for syllogistic reasoning (2015)

Harrop: A new tool in the kitchen of intuitionistic logic

Andrea Condoluci[1] and Matteo Manighetti[2]

[1] Department of Computer Science and Engineering, Università di Bologna, Italy
[2] INRIA Saclay & LIX, École Polytechnique, Palaiseau, France

Abstract. The usual reading of logical implication $\varphi \to \psi$ as "if φ then ψ" fails in intuitionistic logic: there are formulas φ and ψ such that $\varphi \to \psi$ is not provable, even though ψ is provable whenever φ is provable. Intuitionistic rules apparently cannot derive interesting meta-properties of the logic and, from a computational perspective, the programs corresponding to intuitionistic proofs are not powerful enough.
We begin our quest for a better computational understanding of the intuitionistic connectives from *Harrop's rule*, which we consider in this paper. By adding its corresponding axiom $(\neg p \to q \vee r) \to (\neg p \to q) \vee (\neg p \to r)$ to propositional logic, one obtains the Kreisel-Putnam logic **KP**: we give a Curry-Howard correspondence for this system, proving the disjunction property and all the good constructive properties.
This is a first step in understanding how the programs of admissible rules look like.

Keywords: intuitionistic logic, admissible rules, Harrop's rule, Curry-Howard correspondence

1 Introduction

Axiomatic proof systems are presented by giving axioms and rules of inference, which are respectively the ingredients and the tools for cooking new proofs. For example, when presenting *classical propositional logic* (**CPC**) in *natural deduction*, for each of the usual connectives $\wedge, \vee, \neg, \to, \bot$ one gives a set of standard tools to introduce or remove that connective from a formula in order to obtain a proof.

Rules have the form $\varphi_1, \ldots, \varphi_n / \psi$ (read "from $\varphi_1, \ldots, \varphi_n$ infer ψ") where $\varphi_1, \ldots, \varphi_n, \psi$ are schemata of logic formulas. A rule is admissible in a proof system if it is in a way redundant:

Definition 1 (Admissible rule). *A rule $\varphi_1, \ldots, \varphi_n / \psi$ is said to be* admissible *in a system X if whenever $\varphi_1, \ldots, \varphi_n$ are provable then ψ is provable. More formally:*

$$\text{if } \vdash_X \varphi_1 \ldots \vdash_X \varphi_n \text{ then } \vdash_X \psi.$$

Clearly a rule of inference which is already provided by a proof system is also admissible in that system, but this case is not interesting: surely one can use any tool their kitchen provides. But adding or dropping rules may increase or decrease the amount of proofs we can cook in a proof system. The effect can be dramatic: for example, classical logic can be obtained by simply adding the rule of *double negation elimination* to *intuitionistic propositional logic* (**IPC**). Admissible rules are all the opposite: if we decide to utilize one in order to cook something, then we could have just done things in a different way and obtain the same result.

One appealing feature of **CPC** is being *structurally complete*: all admissible rules are *derivable*, in the sense that whenever $\varphi_1, \ldots, \varphi_n/\psi$ is admissible, then also $\varphi_1 \wedge \cdots \wedge \varphi_n \to \psi$ is provable [3] – *i.e.* the system acknowledges that there's no need for that additional tool. This is not the case in intuitionistic logic: the mere fact that we *know* that the tool was not needed, doesn't give us any way to show inside the system *why* is that. On the other hand, **IPC** has other wonderful features. Relevant here is the *disjunction property*, fundamental for a constructive system: when a disjunction $\varphi \vee \psi$ is provable, then one of the disjuncts φ or ψ is provable as well.

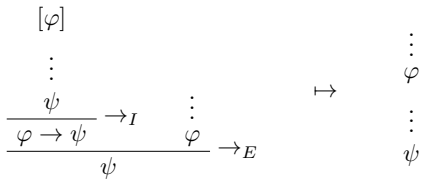
Our interest is in the intuitionistic admissible rules that are not derivable, and the logic systems obtained by adding explicitly such rules to **IPC**. A crucial but non-trivial question is whether admissible rules may break the constructive nature of the logic, and interestingly we are going to prove that the disjunction property still holds in our case.

Can one effectively identify all intuitionistic admissible rules? The question of whether that set of rules is recursively enumerable was posed by Friedman in 1975, and answered positively by Rybakov in 1984. It was then de Jongh and Visser who exhibited a numerable set of rules (now known as *Visser's rules*) and conjectured that it formed a basis for all the admissible rules of **IPC**. This conjecture was later proved by Iemhoff in the fundamental [5]. Rozière in his Ph.D. thesis [7] reached the same conclusion with a substantially different technique, independently of Visser and Iemhoff. These works elegantly settled the problem of identifying and building admissible rules. However our question is different: *why* are these rules superfluous, and what reduction steps can eliminate them from proofs?

In sequent calculus, the most iconic admissible rule is the *cut rule*: yes, one can avoid chopping in intuitionistic logic and Gentzen showed how. His cut-elimination algorithm removes redundant cut inferences, obtaining cut-free proofs and showing that the cut is unnecessary. By the Curry-Howard correspondence, one may assign proof terms to proofs in such a way that removing cuts from proofs corresponds to normalizing proof terms according to some reduction rules. Let's apply this intuition to the admissible cookware: what proof transformation/reduction rules correspond to other admissibile rules?

First of all, note that standard reductions in natural deduction correspond to removing unnecessary *detours*, *i.e.* introduction inferences followed by elimi-

nation inferences. For example, β-reduction in lambda-calculus corresponds in natural deduction to eliminating a $\to_I$ followed by a $\to_E$:



Reduction: $(\lambda x.\, u)\, t \quad \mapsto \quad u\{t/x\}$

Our plan is to understand the phenomenon of admissibility by equipping proofs with lambda terms and associated reductions in the spirit of the example above. The detour removal procedure will show explicitly what is the role that admissible rules play in a proof. In this work, we concentrate our effort on Harrop's rule, historically the first admissible rule, and then we propose how to generalize the technique to harder cases.

1.1 Outline of the paper

In Section 2, we introduce the Kreisel-Putnam logic by adding Harrop's principle to intuitionistic logic; we give a Curry-Howard correspondence for the calculus, providing a computational interpretation of that principle admissible rule. In Section 3 we discuss some usual properties of type systems, which we use to prove the Disjunction property (Theorem 3). We conclude in Section 4 by presenting future extensions to bigger systems.

2 The Kreisel-Putnam logic KP

The formulas of propositional intuitionistic logic are defined inductively from propositional atoms and logical connectives. We use greek letters like $\alpha, \beta, \gamma, \varphi, \psi$ to denote formulas. The negation is defined as $\neg\varphi := \varphi \to \bot$.

Harrop's rule [3] was the first rule to be shown to be admissible but not derivable in intuitionistic logic. It consist of the following rule of inference

$$(\neg\psi \to \alpha \vee \beta) \;\;/\;\; (\neg\psi \to \alpha) \vee (\neg\psi \to \beta)$$

and can be seen as the simplest case arising from Visser's basis of admissible rules. The logic **KP** is obtained by adding the corresponding principle (*i.e.* forcing the rule to be derivable) to **IPC**, and was introduced by G. Kreisel and H. Putnam in [6] to exhibit a logic stronger than **IPC** that preserves the disjunction property, thus disproving the conjecture of Łukasiewicz that no such logic could exist.

We present **KP** in natural deduction: it consists of all inference rules of **IPC** $(ax, \wedge_I, \wedge^i_E, \vee^i_I, \vee_E, \to_I, \to_E, \bot_E)$ plus Harrop's rule in the following form:

$t, u, v ::= x$		(variable)	Evaluation contexts for **IPC**:
	$\mid u\, v$	(application)	$C ::= [\cdot]$
	$\mid \lambda x.\, t$	(abstraction)	$\mid C\, t \mid \mathtt{efq}\, C \mid \mathtt{proj}_i\, C$
	$\mid \mathtt{efq}\, t$	(exfalso)	$\mid \mathtt{case}[C \parallel - \mid -]$
	$\mid \langle u, v \rangle$	(pair)	
	$\mid \mathtt{proj}_i\, t$	(projection)	Evaluation contexts for **KP**:
	$\mid \mathtt{inj}_i\, t$	(injection)	$E ::= [\cdot]$
	$\mid \mathtt{case}[t \parallel y.u \mid y.v]$	(case)	$\mid C[E]$
	$\mid \mathtt{hop}[x.t \parallel y.u \mid y.v]$	(Harrop)	$\mid \mathtt{hop}[x.E \parallel - \mid -]$

Fig. 1. Proof terms (left) and evaluation contexts (right)

$$\mathrm{H}:\ \dfrac{\begin{array}{ccc}[\neg\psi]_{(n)} & [\neg\psi \to \alpha]_{(n)} & [\neg\psi \to \beta]_{(n)} \\ \vdots & \vdots & \vdots \\ \alpha \vee \beta & \varphi & \varphi\end{array}}{\varphi}\,(n)$$

Our adaptation of Harrop's rule in natural deduction has the form of an elimination rule (it may be used to derive any formula φ), and instead of requiring the hypothesis $\neg\psi \to \alpha \vee \beta$, we ask for a proof of $\alpha \vee \beta$ with the additional assumption of $\neg\psi$. In this way, the rule looks much alike the usual $\vee_E$: in fact, we can see it as the elimination of a disjunction that may have an additional negative hypothesis.

Now, clearly Harrop's rule is derivable in **KP**:

$$\dfrac{\dfrac{[\neg\psi \to \alpha \vee \beta]_{(2)} \qquad [\neg\psi]_{(1)}}{\alpha \vee \beta} \qquad \dfrac{[\neg\psi \to \alpha]_{(1)}}{(\neg\psi \to \alpha) \vee (\neg\psi \to \beta)} \qquad \dfrac{[\neg\psi \to \beta]_{(1)}}{(\neg\psi \to \alpha) \vee (\neg\psi \to \beta)}\,(1)}{\dfrac{(\neg\psi \to \alpha) \vee (\neg\psi \to \beta)}{(\neg\psi \to \alpha \vee \beta) \to (\neg\psi \to \alpha) \vee (\neg\psi \to \beta)}\,(2)}$$

We assign *proof terms* to proofs in **KP** according to the Curry-Howard tradition: the language of terms is shown in Figure 1 on the left. The proof term annotations for the usual intuitionistic rules are standard (see for example [8]); the annotation for the rule H is:

$$\dfrac{\Gamma, x\colon \neg\psi \vdash t\colon \alpha \vee \beta \qquad \Gamma, y\colon \neg\psi \to \alpha \vdash u_1\colon \varphi \qquad \Gamma, y\colon \neg\psi \to \beta \vdash u_2\colon \varphi}{\Gamma \vdash \mathtt{hop}[x.t \parallel y.u_1 \mid y.u_2]\colon \varphi}$$

As expected, it is reminiscent of the `case` annotation, with the difference that it also binds the variable x in the first entry t. The other terms bind in the usual way: $\lambda x.\, t$ binds x, $\mathtt{case}[t \parallel y.u_1 \mid y.u_2]$ binds y in u_1 and u_2, and $\mathtt{hop}[x.t \parallel y.u_1 \mid y.u_2]$ binds x in t and y in u_1, u_2.

The reduction rules for the proof terms are given in Figure 2: once again the first block contains the usual ones for **IPC** (with the slightly non-standard *Exfalso* rule that pushes uses of $\perp_E$ to the outermost possible level), and the

– Rules for **IPC**:
  - *Beta:* $(\lambda x.\, t)\, u \mapsto t\{u/x\}$
  - *Exfalso:* $C[\mathtt{efq}\, t] \mapsto \mathtt{efq}\, t$
  - *Projection:* $\mathtt{proj}_i\, \langle t_1, t_2 \rangle \mapsto t_i$
  - *Case:* $\mathtt{case}[\mathtt{inj}_i\, t \parallel y.u_1 \mid y.u_2] \mapsto u_i\{t/y\}$
– Additional rules for **KP**:
  - *Harrop-inj:* $\mathtt{hop}[x.\mathtt{inj}_i\, t \parallel y.u_1 \mid y.u_2] \mapsto u_i\{\lambda x.\, t/y\}$
  - *Harrop-efq:* $\mathtt{hop}[x.\mathtt{efq}\, t \parallel y.u_1 \mid y.u_2] \mapsto u_i\{(\lambda x.\, \mathtt{efq}\, t)/y\}$

Fig. 2. Reduction rules

second block contains reduction rules for the new $\mathtt{hop}[x.t \parallel y.u_1 \mid y.u_2]$ construct, depending on two shapes that t might have. Let us explain the intuition. In the first case, the term is the injection $\mathtt{inj}_i\, t$ with possibly a free variable x of type $\neg\psi$; in that branch one has clearly chosen to prove one of the two disjuncts α or β; we may just reduce to the corresponding proof u_i, in which we plug the proof t but after binding the free variable x. In the second case, the term is an exfalso, *i.e.* the proof uses a contradiction to prove the disjunction. We may reduce to either one of the two branches u_i, now using the contradiction to derive either α or β, of course under the hypothesis $\neg\psi$.

Finally, let's turn to reduction. Contexts are defined intuitively as proof terms with a hole; the hole is denoted by $[\cdot]$, and $E[t]$ means replacing the hole in the context E with the term t. Reduction contexts are defined by the grammar in Figure 1 on the right. Reduction is obtained as usual from reduction rules $\mapsto$ as the contextual closure under evaluation contexts: if $t \mapsto u$ then $E[t] \to E[u]$. We chose carefully the evaluation contexts in order to simplify normal forms: instead of full reduction, we use *weak head reduction*, *i.e.* reductions are performed only in the head of terms, and not under abstractions.

3 Theorems

The calculus presented in the previous section introduced slight variations w.r.t. the standard lambda calculus associated with **IPC**. In this section we show that the calculus indeed preserves the usual properties of *subject reduction* (the reductions preserve the type of a term) and *normalization* (all typable terms reduce to a normal form). We recall that a term is in normal form (in short, *n.f.*) when it cannot be reduced further. We are going to classify normal forms, and finally show the disjuction property. As a side note, the calculus is not confluent (because of the too permissive Exfalso rule and the non-determinism of Harrop-efq): this is irrelevant for our results, but will be fixed in a future version of the calculus.

First of all, two elementary results that we need below. The first one is *substitutivity*, *i.e.* that substitution preserves types: if $\Gamma, x\colon \psi \vdash t\colon \varphi$ and $\Gamma \vdash u\colon \psi$, then $\Gamma \vdash t\{u/x\}\colon \varphi$ (can be proved by induction on the structure of t). The

second one is *inversion*, that allows to read typing deductions backwards, guided by the syntax of terms (can be proved by induction on the type derivation).

Theorem 1 (Subject reduction). *If $\Gamma \vdash t : \varphi$ and $t \to t'$, then $\Gamma \vdash t' : \varphi$.*

Proof. By the definition of reduction as the closure of $\mapsto$ under evaluation contexts, we just prove the statement when $t \mapsto t'$; the general case $t \to t'$ follows because substitution preserves types.

The cases of the usual intuitionistic reductions are standard (see for example [8]); we just prove the cases of the reduction rules associated with `hop`.

For the case of the left injection $\texttt{hop[}x\texttt{.inj}_1\, t \parallel y\texttt{.}u_1 \mid y\texttt{.}u_2\texttt{]} \mapsto u_1\{\lambda x.\, t/y\}$, by inversion we have $\Gamma, y : \neg\psi \to \alpha \vdash u_1 : \varphi$ and $\Gamma, x : \neg\psi \vdash \texttt{inj}_1\, t : \alpha \vee \beta$ for some $\alpha, \beta, \psi, \varphi$. Again by inversion $\Gamma, x : \neg\psi \vdash t : \alpha$, and by $\to_I$ we obtain $\Gamma \vdash \lambda x.\, t : \neg\psi \to \alpha$ By substitutivity we get the desired result $\Gamma \vdash u_1\{\lambda x.\, t/y\} : \varphi$. The case of the right injection is analogous.

Finally, if $\texttt{hop[}x\texttt{.efq}\, t \parallel y\texttt{.}u_1 \mid y\texttt{.}u_2\texttt{]} \mapsto u_1\{\lambda x.\, \texttt{efq}\, t/y\}$, by inversion we have $\Gamma, y : \neg\psi \to \alpha \vdash u_1 : \varphi$ and $\Gamma, x : \neg\psi \vdash \texttt{efq}\, t : \alpha \vee \beta$ for some $\alpha, \beta, \psi, \varphi$. Again by inversion $\Gamma, x : \neg\psi \vdash t : \bot$, and by $\bot_E$ we obtain $\Gamma, x : \neg\psi \vdash \texttt{efq}\, t : \alpha$. By $\to_I$ we obtain $\Gamma \vdash \lambda x.\, \texttt{efq}\, t : \neg\psi \to \alpha$, and by substitutivity we get the desired result $\Gamma \vdash u_1\{\lambda x.\, \texttt{efq}\, t/y\} : \varphi$.

Theorem 2 (Strong normalization). ***KP*** *is* strongly normalizing.

Proof. A proof of normalization is obviously too large for this short paper, but we still want to give a sketch of the proof. One can use the reducibility method (see for example [8]), among the most powerful techniques for proving normalization up to higher-order logic. The usual proof for the simply typed lambda calculus can be extended to product and sum types (*i.e.* conjunctions and disjunctions) and the type of absurdity (*i.e.* $\bot$). For **KP**, one just extends the definition of *inert terms* to include Harrop terms. The additional reductions for `hop` are similar to the reductions for `case`, and should not give particular problems. We leave a precise proof of normalization for future work.

Before we turn to the *disjunction property* (Theorem 3), we first prove the *Classification* lemma (Lemma 1) and inspect normal forms. While usually the statement of the lemma talks about empty environments ($\Gamma = \emptyset$), we use non-empty ones, but only with negative assumption, *i.e.* contexts of the form $\Gamma_\neg := \{x_1 : \neg\psi_1, \ldots, x_n : \neg\psi_n\}$. Let us explain why: when reasoning on normal forms, one needs to enter recursively the first arguments of Harrop terms, hence going under binders. Therefore the inductive case needs to consider normal forms with additional free variables of negative types, bound by Harrop terms in outer parts of the current term.

Lemma 1 (Classification). *Let $\Gamma_\neg \vdash t : \tau$ for t in n.f. and t not an exfalso:*

- Implication*: if $\tau = \varphi \to \psi$, then t is an abstraction or a variable in $\Gamma_\neg$;*
- Disjunction*: if $\tau = \varphi \vee \psi$, then t is an injection;*
- Conjunction*: if $\tau = \varphi \wedge \psi$, then t is a pair;*

– Falsity*: if $\tau = \bot$, then $t = x\,v$ for some v and some $x \in \Gamma_\neg$;*

Proof. By induction on the type derivation of t:

– (ax) t is a variable in $\Gamma_\neg$. By definition of $\Gamma_\neg$, the type of t is an implication, and we conclude.
– $(\to_I)$ t is an abstraction, and we conclude.
– $(\to_E)$ $t = u\,v$ with $\Gamma_\neg \vdash u\colon \varphi \to \psi$. Because t is in normal form, u cannot be an abstraction or an exfalso. By *i.h.*, u is a variable in $\Gamma_\neg$: hence the type of t is $\bot$, and we conclude because $t = x\,v$.
– $(\vee_I)$ t is an injection, and we conclude.
– $(\vee_E)$ not possible. Assume $t = \mathtt{case}[u \parallel - \mid -]$ with $\Gamma_\neg \vdash u\colon \alpha \vee \beta$, and derive a contradiction. By *i.h.* u is an injection or an exfalso. This contradicts the hypothesis that t is a normal form.
– $(\wedge_I)$ t is a pair, and we conclude.
– $(\wedge_E)$ not possible. Assume $t = \mathtt{proj}_i\,u$ with $\Gamma_\neg \vdash u\colon \alpha \wedge \beta$, and derive a contradiction. By *i.h.* u is a pair or an exfalso. This contradicts the hypothesis that t is a normal form.
– $(\bot_E)$ not possible by the hypothesis that t is not an exfalso.
– (Harrop) not possible. Assume $t = \mathtt{hop}[x.u \parallel - \mid -]$ with $\Gamma_\neg, x\colon \neg\psi \vdash u\colon \alpha \vee \beta$, and derive a contradiction. By *i.h.* u is an injection or an exfalso. This contradicts the hypothesis that t is a normal form.

Our weak head reduction produces *weak head normal forms*:

$$\begin{aligned} V &::= \lambda x.\,t \mid \mathtt{inj}_i\,t \mid \langle u, v\rangle \\ V_\neg &::= V \mid x \mid x\,t \mid \mathtt{efq}\,(x\,t) \end{aligned}$$

Normal forms in a negative environment $\Gamma_\neg$ correspond to the entry $V_\neg$, while values (*i.e.* closed normal forms) correspond to the entry V (it follows from Lemma 1).

Another easy consequence of Lemma 1 is consistency: $\not\vdash t\colon \bot$ for any t. We can finally prove:

Theorem 3 (Disjunction property). *If $\vdash t\colon \varphi \vee \psi$, then $\vdash t\colon \varphi$ or $\vdash t\colon \psi$.*

Proof. Assume $\vdash t\colon \varphi \vee \psi$ with t in normal form. First note that $t \neq \mathtt{efq}\,u$, because otherwise by inversion $\vdash u\colon \bot$, contradicting consistency. By Lemma 1 (with $\Gamma_\neg = \emptyset$) t is an injection. Conclude by inversion.

4 Conclusions and future work

We have added Harrop's principle to intuitionistic logic and we have given a new computational understanding of the resulting system. Thanks to Classification (Lemma 1) we investigated the normal forms of proofs, and shown that proof terms look exactly like usual intuitionistic terms at the outermost level: as expected, Harrop inferences over closed proofs can always be eliminated, making Harrop's rule admissible.

We believe that our presentation is well-suited to continue the study of admissibility in intuitionistic systems. We conclude with some remarks on future generalizations:

Visser's rules and other intermediate propositional logics Our next step is generalizing to more complex instances of Visser's rules. Visser's rule have been shown to be relevant not only for **IPC**, but also for other intermediate and modal logics [4], and thus our system could be related to other such logics that have been given a Curry-Howard correspondence.

First-order logic In intuitionistic first-order theories, the disjunctive and the existential connectives behave likewise, and often reflect similar properties. The most famous case is the *existential property*, clearly linked to the disjunction property that we tackled above: if we can prove an existentially quantified formula, then we can prove the same formula for a concrete witness. The first-order principle corresponding to Harrop's rule is the *Independence of Premise*:

$$(\neg A \to \exists x.\, B(x)) \to \exists x.\, (\neg A \to B(x))$$

As expected, it is an admissible but not derivable rule of intuitionistic logic, and our framework can be easily extended to handle it as well.

Arithmetic Since its inception with Harrop [3], the motivation for studying admissible rules of **IPC** was to understand arithmetical systems. A famous theorem of de Jongh states that the propositional formulas whose arithmetical instances are provable in *intuitionistic arithmetic* (**HA**) are exactly the theorems of **IPC**, and many studies of the admissible rules of **HA** (like Visser [9], Iemhoff and Artemov [1]) originated from it.

The Independence of Premise has an important status in the theory of arithmetic, and was given a constructive interpretation for example by Gödel [2]. Many other admissible rules of **HA**, such as Markov's rule, have been studied for a long time. Therefore, we believe that a substantial field of application of our technique is the constructive study of admissible rules of **HA**.

References

[1] Sergei N. Artemov and Rosalie Iemhoff. "From de Jonghs theorem to intuitionistic logic of proofs". In: Dick de Jonghs Festschrift. 2004, pp. 1–10.

[2] Kurt Gödel. "Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes". In: *dialectica* 12.3 (1958), pp. 280–287.

[3] Ronald Harrop. "On disjunctions and existential statements in intuitionistic systems of logic". In: *Mathematische Annalen* 132.4 (1956), pp. 347–361.

[4] Rosalie Iemhoff. "Intermediate logics and Visser's rules". In: *Notre Dame Journal of Formal Logic* 46.1 (2005), pp. 65–81.

[5] Rosalie Iemhoff. "On the admissible rules of intuitionistic propositional logic". In: *The Journal of Symbolic Logic* 66.1 (Mar. 2001), pp. 281–294.

[6] Georg Kreisel and Hilary Putnam. "Eine Unableitbarkeitbeweismethode für den intuitionistischen Aussagenkalkül". In: *Archiv für mathematische Logik und Grundlagenforschung* 3.3 (Sept. 1, 1957), pp. 74–78.

[7] Paul Rozière. "Admissible rules and backward derivation in intuitionistic logic". In: *Math. Struct. in Comp. Science* 3.3 (1993), pp. 129–136.

[8] Morten Heine Sørensen and Pawel Urzyczyn. *Lectures on the Curry-Howard isomorphism*. Vol. 149. Elsevier, 2006.

[9] Albert Visser. "Substitutions of Σ_1^0-sentences: explorations between intuitionistic propositional logic and intuitionistic arithmetic". In: *Annals of Pure and Applied Logic*. Troelstra Festschrift 114.1 (Apr. 15, 2002), pp. 227–271.

Simulating the No Alternatives Argument in a Social Setting

Lauren Edlin

Faculty of Sociology, University of Bielefeld, Germany
ledlin@uni-bielefeld.de

Abstract. This paper formalizes and simulates a social version of the No Alternatives Argument (NAA). The Social NAA predicts that strength of belief in the most strongly held hypothesis in a group of agents will increase when the number of available hypotheses decreases. Social network simulations using connected Bayesian networks show that this assumption can be violated, but infrequently. Implications of the Social NAA and when it holds in social networks are discussed.

Keywords: No Alternatives Argument · Bayesian Epistemology · Social Networks.

1 Introduction

This paper offers an initial investigation into how the number of choices perceived as feasible by individual agents in a group setting may influence choice at the group level.

This topic is worth considering for several reasons. First, most formal models of decision-making do not account for changes in the number of choices that agents consider before making their decisions. However, theoretical frameworks (e.g., [9]) and subsequent empirical studies (e.g., [11]) in psychology have shown that the perceived number of viable choices does influence an individual agents choice. Second, this influence has not been investigated in a social context, despite cases where the number of perceived choices is relevant at the group level. One pertinent example of where this might be important is voting behavior in a political election. A candidate that a number of particular individuals would normally consider untenable may over time be regarded as a viable option for a variety of reasons. Such reasons may include the fact the candidate is determined to be the best option in light of the other choices, new information makes the candidate a more attractive choice, or the shared opinions of other individuals in the agents' social community convince the agents to change their minds about the candidates. If enough individual agents begin to consider the candidate a viable option, then that candidate is intuitively likely to gain votes come election time.

The relation between the number of choices at the individual level and the outcome at the group level, however, is not straightforward. On the one hand,

it is difficult to generalize a descriptive account because agents differ in their decision-making processes, which also depend on the context and type of decision being made. On the other hand, it is not intuitively clear what a normative account should capture - questions arise such as whether an agent should take more caution when increasing the number of available options. As a preliminary step in determining reliable relations between individually perceived hypotheses and decision outcomes at the group level, this paper tests an assumption made in a recent paper by [5] that supports an argument for the validity of the "no alternatives argument (from here on, 'NAA') in the framework of Bayesian epistemology (e.g., [4]). Dawid et al. consider the case in which only one hypothesis is regarded as a viable option according to a scientific community, and assert that as the number of alternative hypotheses decreases to zero the strength of belief in the most viable hypothesis increases. A new Bayesian network model that explicitly includes sharing opinions between agents is developed to test this assumption in a social setting through several simulations over various models.

The structure of the rest of the paper is as follows: Section 2 describes Dawid et al.'s formal argument for the NAA in detail. Section 3 develops the new Bayesian network model in light of criticisms of the formal version of the NAA. Section 4 outlines the parameters, including the number and configurations agents in a social network, and the results of several simulations over several versions of the model. Section 5 concludes with a discussion of the results and what they entail for future work on this topic.

2 Dawid et al.'s NAA

The NAA is the argument that, given a hypothesis H, in cases where there has been considerable effort to find an alternative to H and none can be found, the one hypothesis that has been found is more likely to be correct and it is rational for one to strengthen their belief in H. Dawid et al. develop a Bayesian network to prove that an increase in degree of belief in the proposition T, that H is an empirically adequate hypothesis, given the non-empirical evidence F_A, that a community has not found an adequate alternative to H, is greater than T alone: $P(T \mid F_A) > P(T)$. T and F_A are both binary variables, meaning that T represents whether H is believed to be empirically adequate or not, and F_A represents whether a scientific community has found more than one hypothesis for a given phenomenon or set of data.

Since F_A is unable to act as justification for T because it is not logically or probabilistically within the domain of H, indirect factors that mediate the influence of F_A on T are necessary to consider. One factor is the number of alternative hypotheses as determined by the community, which is expressed in the variable Y, and directly influences both F_A and T. In addition, the cumulative factors that determine the difficulty of developing alternative hypotheses, such as the available evidence and the configuration and cleverness of the community, D, also directly influences F_A. Both Y and D take the set of natural numbers as their values.

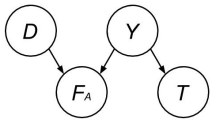


Fig. 1. Bayesian network representation of the NAA

The conditional distributions of the mediating factors Y and D are constructed to reflect the intuition that if there are a large number of alternative hypotheses, then it is likely a community would have found at least some of them as long as the complexities of the problem are not exorbitant. More precisely, Y and D influence F_A independently - for every fixed level of difficulty of the scientific problem at hand, an increase in the number of alternative hypotheses does not decrease the likelihood that scientists have found an alternative hypothesis, and for a fixed number of alternative hypotheses, the increase in the difficulty of the problem does not increase the likelihood of finding an alternative hypothesis. Y also influences T in that an increase in alternative hypotheses does not make it more likely that scientists have found an empirically adequate hypothesis. Figure **??** represents the resultant Bayesian network.

While Dawid et al.'s formal argument for the NAA has faced some criticism ([8]; [2]), the explicit inclusion of the perceived number of available hypotheses in an epistemic model is a novel contribution. Despite this innovation, it is unclear whether Y is meant to describe the beliefs regarding the number of hypotheses of individuals within the scientific community, or the beliefs of the scientific community as a whole. Furthermore, the development of hypotheses is a community effort and in large part relies on communication of information, opinions and knowledge between individuals in a community.

The intuition behind this formalized version of the NAA, namely that the degree of belief in H increases as the number of alternative hypotheses decreases, is now tested in an explicitly social setting in order to account for some of the social processes involved in hypothesis formation.

3 New Model: Agents in a Social Setting

In response to the previous criticisms, a Bayesian network is devised to describe the influence of shared opinions on individual agents beliefs regarding the number of feasible hypotheses. A community network is built through repetitions of a Bayesian network structure that represents an individual agent, which are connected in a way to represent opinion sharing between specified agents.

It is worth noting that numerous models have been developed to investigate information and knowledge dynamics through communications of agents in a community. Models using dynamic epistemic logic, for instance, are used to model the phenomena of information cascades and pluralistic ignorance (e.g., [1]), while probabilistic (e.g., [6]) and simulation models (e.g., [7]) have shown

group dynamics of consensus and polarization, among other social processes based on opinion sharing between agents. Despite the success of these models, the model developed here continues with the Bayesian framework employed by Dawid et al. This is in part to test their assumption within their own framework, and also because multiple hypotheses can easily be accommodated in Bayesian networks. The extension of previously developed models to further investigate the effects of changes in the number of perceived hypotheses at the group level is sanctioned for future work.

The model is described in non-technical language, and begins with defining individual agents before scaling up to a community network.

3.1 Individual Agent Networks

The categorical variable H represents the set of hypotheses under consideration by the agent. If three hypotheses are under consideration, for instance, then the values of H are the set $\{H1, H2, H3\}$. The probabilities assigned to the values constitute a distribution and therefore sum to 1.

Two types of priors determine the distribution in H, which are represented by the E variables and the O variables. The E variables (with "E" standing for "experience") represent the credence afforded to each hypothesis according an agents own experiences, experimentations, values and biases up to a specific point in time. The O variables (with "O" standing for "others") represent the credence afforded to each hypothesis according to the collated opinions of other agents in the network.

The number of E and O variables reflects the number of hypotheses under consideration. Keeping with the example of three hypotheses in H, the network would include three E variables and three O variables, which will be respectively labeled E_H1, E_H2, and E_H3 for the E variables, and O_H1, O_H2 and O_H3 for the O variables.

Each E and O variable is a categorical, binary variable. For each variable, the value 0 (representing $\neg E\_Hn$ and $\neg O\_Hn$ for a particular E and O variable) is interpreted to mean the hypothesis indicated by that variable is considered an unfeasible option for whatever reason (e.g., it is not supported with sufficient evidence, it is not well known to the agent, there is believed to be a low chance that this option will have an impact, etc.). The value 1 (representing E_Hn or O_Hn for a particular E and O variable) is interpreted to mean the hypothesis indicated is perceived to be a feasible option. For each variable, for a given proposition E_Hn (or O_Hn) it will always be the case such that $P(E\_Hn) + P(\neg E\_Hn) = 1$ (or $P(O\_Hn) + P(\neg O\_Hn) = 1$) in line with classical probability theory, and subsequently Bayesian networks.

The E and O variables are priors that define the posterior probabilities in H. The network structure, which is uniform for all agents, is shown in Figure 2. This figure continues with the example of three hypotheses, and represents an individuals belief states the influence the hypotheses in H.

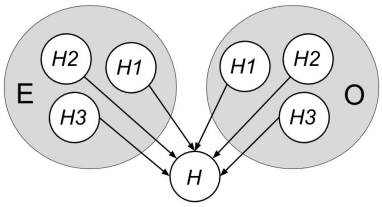


Fig. 2. A Bayesian network representing the belief states of one individual agent considering three hypotheses

3.2 Conditional Probability Assignments Weighting Self and Other Beliefs

Agents take into account the opinions of others and bias their own preconceived beliefs to various extents. Some agents, for instance, may readily take on board the opinions of others in cases where they are believe themselves to be less informed or consider others more expert, while other agents may consistently disregard the opinions of others and stick to their own experiences to support their beliefs. Such differences between agents are represented through variances in the conditional probabilities of the E and O variables on H.

The conditional probabilities of the E and O variables on H for each specified agent can be assigned categorically or randomly. Categorical assignments refer to three types of agents determined by the specified weights on the E and O variables in H. In this paper, three types of agents are defined, respectively labeled as *balanced*, *mule*, and *sheep* type agents. A balanced agent gives equal credibility to its own preconceived opinions and the opinions of others, and therefore each E variable is given a weight of 0.5 in H, and each O variable is given a weight of 0.5 in H. A mule agent gives greater credibility to its own preconceived opinions, and therefore gives each E variable a weight of 0.9 in H and each O variable a weight of 0.1 in H. A mule type agent is consequently much less likely to change its preferred hypothesis in light of its neighbor's opinions. Finally, a sheep agent is weighted in the converse way to the mule, with a weight of 0.1 given to each E variable in H and a weight of 0.9 given to each O variable in H. A sheep type agent is therefore more likely to change its preferred hypothesis in light of the opinions of its neighbors.

Weights defining the E and O variables conditional on H can also be randomly assigned, allowing for more fine-grained conceptions of the extent that an agent may take others opinions on board. However, it is beneficial to define agent types categorically in order to track changes in opinion dynamics according to the type of agent and its place in the larger network.

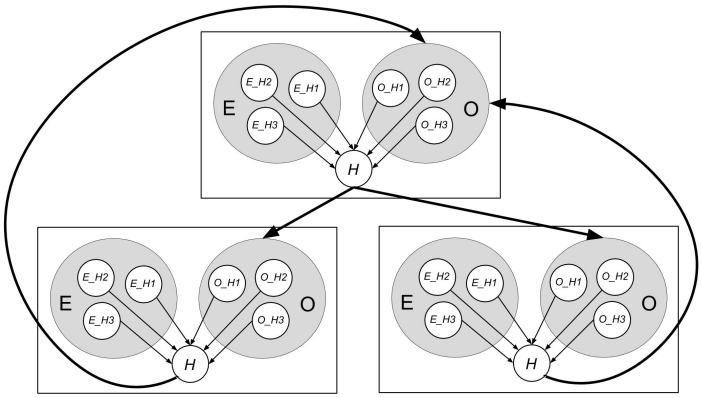


Fig. 3. A three-agent network, with each agent considering three hypotheses

3.3 Scaling up to a Social Network: Collating O Variable Values from an Agents Social Influencing Neighbors

The probabilities of the O variables for agents in a community network are determined by the values in H from specified neighbor agents. More precisely, the hypothesis in H with the highest probability for each agent that is specified to inform or influence a neighbor agent are collated to determine the values in the influenced agents' O variables.

A *parent* agent refers to an agent that influences another's O variables, while a *child* agent is influenced by a parent agent. Each O variable of an agent aggregates the top H value of its parents, and uses that to set new values for its own O variables. For example, if an agent has three parents and two of those parents preferred $H1$ and one preferred $H2$, the child agent would set its O_H1 variable's distribution as $1 = \frac{2}{3}$ and $0 = \frac{1}{3}$, and set its O_H2 variable's values to $1 = \frac{1}{3}$ and $0 = \frac{2}{3}$. All the other O_Hn variables would have the distribution $1 = 0$ and $0 = 1$ (i.e. the other hypotheses have 0 probability according to the agent's parents).

Figure 3 depicts a three-agent network, in which the top agent is the parent to the two lower child agents, influencing their belief through their O variables. As the recursive loop from the children's H to their parents shows that, the child agents also inform the original parent agent's O variables. However, since Bayesian networks are by definition Direct Acyclic Graphs that forbid looping configurations, the influence of the two lower agents on the top agent must be done in a a subsequent belief update step, after they have been influenced by their parent. The update procedures for the simulations are explicated in the following section.

4 Simulations to test the Social NAA

A series of simulations are conducted over various configurations of social networks to test whether the NAA assumption holds in all cases on a social level. What will be termed the *Social NAA* is the hypothesis that as the total number of hypotheses available in the social network (i.e. the number of different hypotheses outputted as a top hypotheses by the agents in their respective H posterior distributions) decreases, the strength in belief in the hypothesis H_{max}, the hypothesis chosen by highest number of agents in the network increases should stay the same or increase– i.e., the same number, or more agents should "vote" for H_{max}, whatever it may be, at each time step until convergence whenever a hypothesis drops out of the possible options with no votes. Networks that violate this outcome are counter-arguments against the social version of the NAA assumption.

For the each simulation, the update procedure for each agent at each time step is constituted by the posterior distribution for H being calculated as follows in two steps:

1. CALCULATE :
   - A **IF** at time-step 0, use the prior distributions for O and E values to calculate the posterior distribution over H using Hs conditional probability table, and calculating the marginal probabilities in the standard way [10].
     ELSE If at time-step 1 or later, update the O variables by aggregating the top hypotheses from the agent's parent's nodes, set these as the priors for O. If this distribution has changed from the previous time-step, calculate the posterior over H.
   - B Add the agent's top hypothesis after calculating the posterior to the overall 'vote' distribution for that time-step, to be recorded for later analysis.
2. UPDATE: Update the agents E variables to be equivalent to that of the posteriors over H calculated at Step 1.

Convergence occurs when agents no longer update their hypothesis belief distribution once computing the posterior H distribution and the distribution from one time-step to the next is identical. In the development of the simulations, it was found a maximum of 10 updates was sufficient to ensure convergence for the particular network configurations used.

All simulations were implemented in Python 2.7 and run using iPython notebooks.

4.1 Simulation 1: Randomized, Categorical Ten Agent Network

For the first simulation, a ten-agent network configuration was adapted from [3] as exemplar of a *core-periphery network* - that is, several highly interconnected agents are in the network core and less connected agents constitute the network

periphery. It serves as an example of a network configuration ubiquitous across a wide variety of communities in reality according to empirical social network literature.

Agents were set to consider four hypotheses in $\{H1, H2, H3, H4\}$, though various stages in the simulation some agents may not be aware of one or more of them (i.e. being assigned a probability of 0).

In this simulation, agents were defined categorically (i.e., as balanced, mule, or sheep agents – see 3.3), and each agent was assigned a category randomly for each run. The E variable priors were also randomized with each run, set to believe one of the four hypotheses and the initial O variable priors were set to a normal distribution over the four hypotheses before being influenced by the beliefs states of other agents. The simulation was run 1000 times with 10 update steps per run.

Figure 4 depicts the configuration of the network, as well as changes in the top hypothesis chosen by each agent at each time step in the run. The numbered boxes refer to agents and their place in the network, and the lines with arrows represent parent and child influences between agents. Each color corresponds to a specific hypothesis, and the color of the agent box indicates the agents top hypothesis (i.e. the hypothesis with the highest probability). This example depicts a successful run demonstrating the social NAA, because the top hypothesis chosen by most of the agents (hypothesis $H4$ in green) increased in strength (i.e. gained more "votes") with each time step as the number of hypotheses considered viable by each agent decreased.

From the 1000 runs, this simulation resulted showed 14 runs violating the Social NAA assumption (i.e. when for some time-step, the strength in the top hypothesis did not increase as the number of hypotheses considered viable by each agent decreased). This supports that this assumption is too strong to describe a general relation between individual hypothesis preferences and outcomes at the group level. Nevertheless, the number of violations was relatively small, which suggests that the social NAA assumption may serve as a useful heuristic on the social level than a rule. No significant difference was found between the distributions of the agent categories (the distribution of mule, sheep and balanced agents) in the simulation runs that violated, and did not violate the Social NAA. More thorough investigation with larger simulations using different categories of agents will be carried out in future.

4.2 Simulation 2: Randomized Twenty Agent Network

In the second simulation, a twenty-agent network was created consisting of two clusters of a core-periphery type networks connected by one agent forming the best connected agent of the second cluster 'bridging' between the two clusters. This again was run 1000 times with 10 simulations per run.

In this simulation, rather than using agent categories, agent types were defined through randomized weights on the conditional probabilities whereby the amount of weight given to an agent's own beliefs (verses its parents) was randomized between $[0, 1]$ for each run. Agents were again set to consider four initial

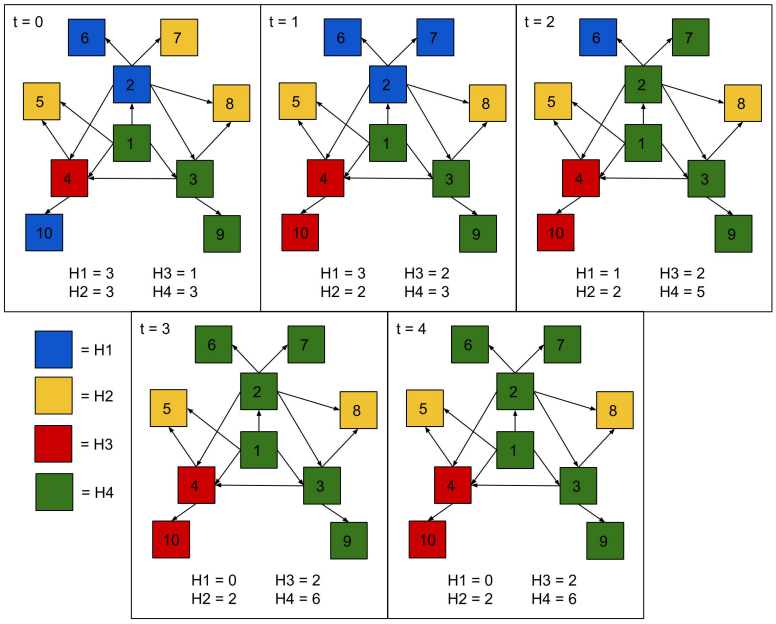


Fig. 4. Example of One Run with Four Updates (First State t = 0). The network converges at time-step 3.

hypotheses. The E variable priors were also randomized with each run, and the initial O variable priors were set to a normal distribution before being influenced by the beliefs states of other agents.

This simulation resulted in 22 violations of the social NAA assumption, which was a similar result to the ten agent network simulation in that a relatively low number of violations occurred. However, the network configurations that violated the social NAA assumption tended to have core agents in the most well-connected position (e.g., agent 1 in Figure 4) which took into account the rest of the network's opinions less than in the configurations which did not violate the social NAA. The average weight of the O variables on H for the best-connected agent in networks that violated the social NAA was 0.376, while the average weight of the O variables on H for best-connected agent in networks that did not violate the social NAA was 0.505. Well-connected agents that did not take others opinions into account would stifle swings to new hypotheses across the network. It was also hypothesized that the violation in this complex network would also depend on the type of the connecting agent between the two clusters, which was effectively a second core agent for the second cluster. It was indeed

found that the strength of weight on other's opinions in non-violating network was higher than in the violating networks– 0.498 vs. 0.300.

5 Conclusions and Future Work

This paper defined a social version of the No Alternatives Argument for belief in hypotheses given the number of hypotheses available to agents in a social network. Violations of the social NAA assumption in simulations support that the assumption is too strong to be part of a valid reasoning process as argued by Dawid et al. Nevertheless, the relatively low number of violations of this assumption is a social context indicates that the social NAA assumption may be a useful heuristic to guide understandings of how individual opinion sharing may influence outcomes of decision-making at the group level.

It must be noted that there are significant shortcomings in this model. First, only a limited number of simulations were conducted. The numerous variables involved in creating the simulations, including configurations of networks, the number of agents, type of agents, and number of hypotheses under consideration, were not fully explored. It is therefore unclear whether the findings from these simulations are just due to the network configurations chosen rather than other variables such as influences based on agent types. While no strong conclusions given the model can be made at this time, further exploration of these parameters is open for future work. Second, the model is only capable of representing uni-directional rather than bi-directional opinion sharing between agents. While the network configurations attempted to account for the lack of bi-directional influences by organizing agents in networks found in empirical social network studies, the full extent of opinion sharing between agents was unable to be explored.

These shortcomings suggest potential for future work. The model in this paper is descriptive in that it attempts to show processes of individual agents decision-making and the outcomes at the group level, rather than argue as to how agents should make decisions in light of group outcomes. One direction for future work is to adapt previously developed models of opinion sharing dynamics (such as those mentioned at the beginning of Section 3) to either account for more realistic descriptions of these processes, or develop a normative account of how agents should employ these processes.

Even more promising is the application of the model to real-world data, as evidence about the actual processes individuals use could then be applied to a normative account. A context that would be well suited for this model is data pertaining to elections. The E variables, which in the current model are under-specified to only provide an overview of preconceived beliefs about the hypotheses under considerations themselves, could be re-conceptualized to capture relevant information about how an agents preconceived preferences are made. One option is to apply the survey data from the American National Election

Studies (ANES) new 2016 Time Series Study.[1] The survey was given before and after the 2016 U.S. General Election, and includes a variety of questions pertaining to political and policy views, opinions about the candidates (such as a "feeling thermometer" on a 0-100 scale, and emotional responses), and previous voting behavior.

In sum, the purpose of this paper is to offer an initial investigation into how the number of choices perceived as feasible by individual agents in a group setting may influence choice at the group level. These findings pertaining to the testing of one assumption about this relation show that this is an area worth considering in future research.

References

1. Baltag, A., Christoff, Z., Hansen, J.U., Smets, S.: Logical models of informational cascades. Studies in Logic **47**, 405–432 (2013)
2. van Basshuysen, P.: Dawid et al.s [2015] no alternatives argument: an empiricist note. Kriterion: Journal of Philosophy **29**(1), 37–50 (2015)
3. Borgatti, S.P., Everett, M.G.: Models of core/periphery structures. Social networks **21**(4), 375–395 (2000)
4. Bovens, L., Hartmann, S.: Bayesian epistemology. Oxford University Press on Demand (2003)
5. Dawid, R., Hartmann, S., Sprenger, J.: The no alternatives argument. The British Journal for the Philosophy of Science **66**(1), 213–234 (2015)
6. DeGroot, M.H.: Reaching a consensus. Journal of the American Statistical Association **69**(345), 118–121 (1974)
7. Douven, I., Riegler, A.: Extending the hegselmann–krause model i. Logic Journal of IGPL **18**(2), 323–335 (2009)
8. Herzberg, F.: A note on the no alternatives argument by richard dawid, stephan hartmann and jan sprenger. European Journal for Philosophy of Science **4**(3), 375–384 (2014)
9. Potter, R.E., Beach, L.R.: Decision making when the acceptable options become unavailable. Organizational Behavior and Human Decision Processes **57**(3), 468–483 (1994)
10. Russell, S.J., Norvig, P.: Artificial intelligence: a modern approach (4th Edition) (2010)
11. Shah, A.M., Wolford, G.: Buying behavior as a function of parametric variation of number of choices. Psychological Science **18**(5), 369 (2007)

[1] `http://www.electionstudies.org/studypages/anes_timeseries_2016/anes_timeseries_2016.htm`

Explainability of Irrational Argument Labelings

Grzegorz Lisowski

ILLC, Universiteit van Amsterdam

Abstract. In this paper we study the problem of providing a justification for a seemingly irrational choice of arguments. The study is set in a framework of value-based argumentation, in which the abstract argumentation is extended with an assignment of values to arguments. These are used to determine the relative strength of arguments. We use this approach to provide methods of generating plausible explanations for argument selection not satisfying certain rationality constraints.

1 Introduction

Abstract argumentation aims at providing explanation for the choice of accepted pieces of information, taking into account conflicts between data available to a decision-maker. Within this framework several constraints regarding the sets of arguments which can be selected as an outcome of the deliberation process have been provided.

However, it might be the case that an agent decides to make a seemingly irrational decision about the selection of arguments. Imagine a situation in which an agent says that she needs to have a cup of coffee at midnight, because she has a paper due the following day. Her friend tells her, however, that she should not drink coffee at night because it is unhealthy. Then, the agent decides to drink coffee even though she agrees that it ruins her health.

In abstract argumentation, as defined by Dung (1995), the decision of whether to accept or not a set of arguments depends only on the attack relationship between them. However, it might be the case that particular arguments have different strengths from different agents' perspectives. If we would only take attack relation between arguments into account while assessing their acceptability, the previously described agent would not be rational. She refused to listen to friend's advice, even though she did not have any counterarguments for her claims. It can be the case, however, that an agent discriminates between the *strength* of available arguments. Then, an agent might be willing to disregard an attack on a strong argument by a weaker one. In the described case, she could have decided to treat an argument stating that she needs to drink the coffee as particularly strong because academic success is the most important for her.

Following a recent trend in the literature on abstract argumentation (e.g. Fan & Toni, 2015), this paper aims at capturing explainability of decisions about the selection of arguments. In particular, we are concerned with the possibility of explanation of seemingly irrational choices of selected arguments by providing assumptions regarding their relative strength from the perspective of a particular

agent. In this way we might detect agents' perceived strength of arguments, under the assumption that they act rationally.

One of the particular ways of determining the strength of arguments for a particular assessor is the value-based argumentation. This approach, due to Bench-Capon (2003), is based on the assumption that a persuasive power of arguments relies in a substantial part on the values an agent assigns to them. Further, an individual view on importance of values influences the agent's view on whether a particular counterargument is strong enough to defeat an argument it is in conflict with. Recently, this approach has been used to explain the differences between agents' perceived conflicts between arguments (Airiau, Bonzon, Endriss, Maudet, & Rossit, 2016).

In our setting it is studied under which circumstances is it possible to map a set of values to an argumentation framework and a preference ordering over them, under which the choice of accepted arguments is justified with respect to a certain rationality constraint.

In Section 2 we provide an introduction of abstract argumentation and Value-Based Argumentation Frameworks. We also define rationality constraints on argument acceptance. Further, in Section 3, we provide results about possibility of successful explainability of argument acceptance with respect to these constraints. Section 4 includes conclusions and directions for further research.

Proofs are omitted, as they involve standard techniques only.

2 The Model

In the abstract argumentation approach, as introduced by Dung (1995), a deliberative process is captured as a set of arguments linked with a binary relation representing the attacks between arguments. The relations between arguments are further used to establish the rationality criteria for selection of accepted arguments.

Definition 1 (Argumentation Framework). *An argumentation framework is a tuple $AF = \langle A, \rightarrow \rangle$, where A is a non-empty set of arguments and $\rightarrow \subseteq A^2$ is a binary attack relation.*

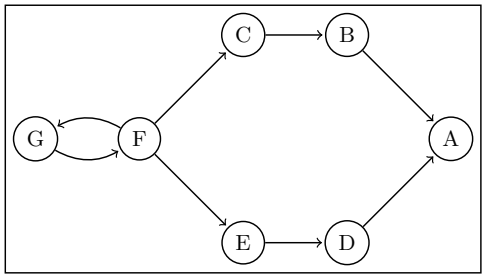


Fig. 1. Example of an argumentation framework.

One of the methods of representing the decisions about acceptance of arguments involves a direct assignment of labels to arguments, denoting their degree of acceptance. It differs from the approach in which just a set of selected arguments is specified. On the contrary, the labeling approach allows not only for capturing a binary decision about argument selection, but also for expressing that a decision-maker is undecided about the status of some argument.

Formally, an *argumentation labeling* is a function from the set of arguments to the set of available labels. We are considering three labels, one denoting firm acceptance (IN), one firm rejection (OUT) and one corresponding to the lack of decision ($UNDEC$).

Definition 2 (Argumentation Labeling). *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework. A labeling over AF is a mapping*

$$\mathcal{L}ab : A \rightarrow \{IN, OUT, UNDEC\}$$

Given $L \in \{IN, OUT, UNDEC\}$, $L(\mathcal{L}ab)$ denotes the set of arguments $\{a \in A | \mathcal{L}ab(a) = L\}$.

Figure 2 illustrates the definition of an argumentation labeling.

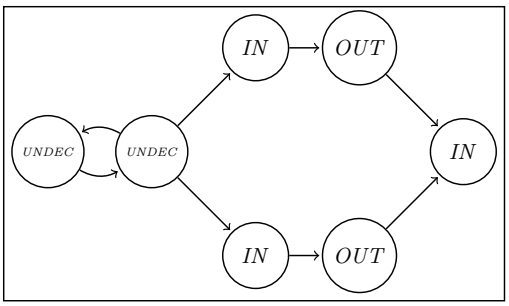


Fig. 2. Example of a labeled argumentation framework.

In order to capture the rationality constraints on the labeling of arguments, it is useful to specify when an argument is labeled in a specific manner legally.

1. It is stipulated that an argument is labeled IN legally if all its attackers are labeled OUT. In this way we can ensure that there is no plausible information conflicting any accepted argument.
2. Further, an argument is labeled OUT legally, if there is some argument labeled IN attacking it. This condition ensures that a decision about a firm rejection of an argument is based on highly plausible grounds.
3. Finally, it is rightful to label an argument $UNDEC$ if it has some attackers who are not labeled OUT, but none of them is labeled IN. With this condition

we express that an individual is only justified in leaving an argument as undecided, if she has some evidence against it, but is not certain about its status.

Formally, these are captured as follows:

Definition 3 (Legal Labelings). *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework, $a \in A$ and $\mathcal{L}ab$ be a labeling over AF. We say that:*

- *$\mathcal{L}ab(a) = IN$ legally iff for all $a' \in A$ such that $a' \rightarrow a$, $\mathcal{L}ab(a') = OUT$*
- *$\mathcal{L}ab(a) = OUT$ legally iff there is some $a' \in A$ such that $a' \rightarrow a$ and $\mathcal{L}ab(a') = IN$*
- *$\mathcal{L}ab(a) = UNDEC$ legally iff for all $a' \in A$ such that $a' \rightarrow a$, $\mathcal{L}ab(a') \neq IN$ and there is some $a' \in A$ such that $a' \rightarrow a$ and $\mathcal{L}ab(a') \neq OUT$.*

We can distinguish the following types of labelings using the legality criteria provided earlier:

Definition 4 (Labeling Semantics). *Let $AF\langle A, \rightarrow \rangle$ be an argumentation framework. We say that a labeling $\mathcal{L}ab$ over AF is:*

- ***Conflict-free*** *if there is no pair of arguments $a, b \in A$ such that $a, b \in IN(\mathcal{L}ab)$ and $a \rightarrow b$.*
- ***Admissible*** *if for all $a \in A$, if $\mathcal{L}ab(a) = IN$, then $\mathcal{L}ab(a) = IN$ legally, and if $\mathcal{L}ab(a) = OUT$, then $\mathcal{L}ab(a) = OUT$ legally.*
- ***Complete*** *if $\mathcal{L}ab$ is admissible and for any $a \in A$, if $\mathcal{L}ab(a) = UNDEC$, then $\mathcal{L}ab(a) = UNDEC$ legally.*

If a labeling satisfies conditions of semantics σ, we call it a σ-labeling.

Let us now proceed to defining the *value-based argumentation*. It will allow us for formalizing the notion of strength of arguments, induced by a hierarchy of values that they refer to.

The basic concept of this approach is that of the *value-based argumentation framework*. It is an extension of an argumentation framework. Further, a set of values is considered. This set is subsequently mapped on the set of arguments. Finally, we consider a preference ordering, representing different views on the hierarchy of values.

Definition 5 (Value-Based Argumentation Framework). *A* value-based argumentation framework *(VAF) is a tuple $\langle A, \rightarrow, V, val \rangle$, where:*

- *A is a non-empty set of arguments.*
- *$\rightarrow \subseteq A^2$ is an attack relation.*
- *V is a non-empty set of values.*
- *$val : A \rightarrow V$ is a function assigning values to arguments.*

Given the definition of the *VAF*, we can express the relative strength of an argument for a given audience. We associate audiences with particular views on the hierarchy of values. This term does not presuppose that there are multiple agents within an audience.

Definition 6 (Audience). *Take a* $VAF = \langle A, \rightarrow, V, val \rangle$*. Then, an* audience *is a linear*[1] *ordering over* V*. We denote that a value* v_1 *is more important than* v_2 *for an audience* P *as* $v_1 \succ_P v_2$*.*

We proceed by evaluation of attacks in the argumentation framework underlying a considered *VAF*. We say that an argument a defeats an argument b for a certain audience P if it attacks it and the value carried by b is not higher than the value of the argument a from P's perspetive. In this way the preference orderings over values are used to establish indirectly the relative strength of arguments.

Definition 7 (Defeat). *Take a* $VAF = \langle A, \rightarrow, V, val \rangle$ *and an audience* P*. Then, for a pair of arguments* $a, b \in A$*,* a defeats b *for* P *(*$a \rightarrow^P b$*) iff* $a \rightarrow b$ *an* $val(b) \not\succ_P val(a)$*.*

Using the notion of defeat for an audience we can convert any *VAF* into the argumentation framework in which an attack relation is replaced with the defeat relation for the audience. In such a graph the defeat relation is always a subset of the initial attack relation. It is worthwhile to emphasize that in the currently described framework the relative strength of arguments imposed by the assignment of values and the preference ordering over them is only used to induce the defeat graph. All acceptability conditions described earlier can then be applied to the defeat graph of the *VAF*. Then, acceptability of arguments can be assessed with the hierarchy of values to which they appeal taken into account.

Definition 8 (Defeat Graph). *Given a value based argumentation framework* $VAF = \langle A, \rightarrow, V, val \rangle$ *and an audience* P*, the* defeat graph *of VAF for* P *is an argumentation framework* $AF = \langle A, \rightarrow^P \rangle$*.*

With the employment of the notions introduced before, we can capture what it means for a labeling to be rational. We say that this is the case if we can find an assignment of values and an audience which would ensure that a labeling satisfies a desired rationality constraint.

Then, a labeling of an argumentation framework is rational with respect to some semantics σ if we can find a *VAF* and an audience P such that it is a σ-labeling of the defeat graph of *VAF* based on P.

Definition 9 (Rationality of Argumentation Labelings). *Let* $\mathcal{L}ab$ *be a labeling of an argumentation framework* $AF = \langle A, \rightarrow \rangle$ *and* σ *be a labeling semantics. We say that* $\mathcal{L}ab$ *is rational iff there is a* $VAF = \langle A, \rightarrow, V, val \rangle$ *and an audience* P *such that* $\mathcal{L}ab$ *is a* σ*-labeling of the defeat graph of VAF based on* P*.*

Let us illustrate the notion of rationality of labeling on an example.

Example 1. Take an argumentation framework labeled in the way shown in the Figure 3.

[1] A linear ordering is an antisymmetric, transitive and connex relation.

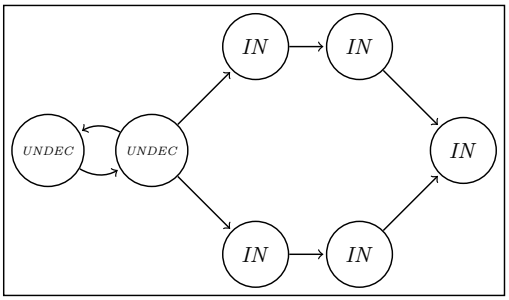


Fig. 3. Example of a non conflict-free argumentation labeling.

Clearly, the conflict-freeness condition is violated in this case, as there are arguments labeled *IN* which are in conflict. This is an indication that the agent submitting this labeling is not rational. However, we can take into account that she is assigning different strength to distinctive arguments. To account for that we might extend this framework to the *VAF* depicted in the Figure 4.

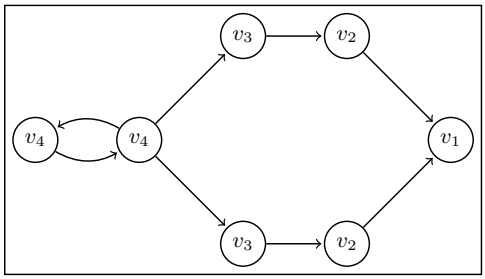


Fig. 4. Example of a *VAF* justifying an irrational labeling.

To establish the relative strength of arguments in this instance, let us consider an audience P:

$$v_1 \succ v_2 \succ v_3 \succ v_4$$

Then, the induced defeat graph of the *VAF* based on P is of the form presented in the Figure 5.

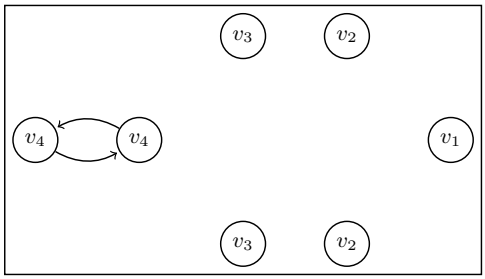


Fig. 5. Defeat graph of the example *VAF*.

Now it is straightforward to check that the labeling presented in the Figure 5 is complete.

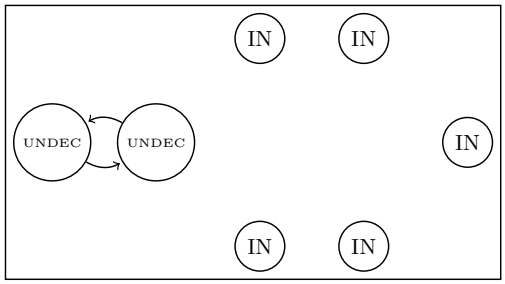


Fig. 6. Labeling of the defeat graph.

So, the initial irrationality of the considered labeling has been explained by the assignment of values and setting a preference ordering over them.

3 Results

In this paper we are concerned with establishing the structural properties of argumentation labelings allowing for their rationality. It is worth noting that different levels of demanded rationality, associated with chosen labeling semantics, might require satisfaction of different requirements. We will investigate them for conflict-freeness, admissibility and completeness conditions.

Let us commence with providing a condition under which a labeling is rational if only conflict-freeness is required.

Proposition 1. *A labeling $\mathcal{L}ab$ is rational with respect to conflict-free semantics iff there are no arguments $a, b \in IN(\mathcal{L}ab)$ s.t. there are paths $a \to \cdots \to b$ and $b \to \cdots \to a$ in which all arguments are labeled IN.*

Let us now extend the previous result to account for admissible semantics.

Proposition 2. *A labeling $\mathcal{L}ab$ is rational with respect to admissible semantics iff $\mathcal{L}ab$ is rational with respect to conflict-free semantics and for any $a \in OUT(\mathcal{L}ab)$ there is some $b \in IN(\mathcal{L}ab)$ s.t. $b \to a$*

Then, we can show a requirement for rationality with respect to complete semantics.

Proposition 3. *A labeling $\mathcal{L}ab$ over $AF = \langle A, \to \rangle$ is rational with respect to complete semantics iff $\mathcal{L}ab$ is rational with respect to admissible semantics and for any $a \in UNDEC(\mathcal{L}ab)$ there is some $b \in UNDEC(\mathcal{L}ab)$ s.t. $b \to a$*

4 Conclusions

In this paper we have have introduced an approach for explainability of seemingly irrational decisions regarding acceptance of arguments, with employment of value-based argumentation frameworks. We used *labeling semantics* as the criteria of rationality. Further, we have studied properties of argumentation labelings which allow for its rationality under conflict-free, admissible and complete semantics.

4.1 Future Work

The results presented in this paper leave room for further research. It would be beneficial to provide explanations of decisions irrational with respect to the current approach with other plausible methods, such as involve adding arguments to the initial framework. This could allow for rationalizing larger classes of argumentation labelings.

Furthermore, it would be of interest to study how to rationalize argumentation labelings under minimal assumptions with respect to agents preferences over arguments. This involves the use of the minimal number of assigned values, or choosing the assignment which allows for blocking as few attacks as necessary.

Finally, it would be interesting to study the computational complexity of checking if a given labeling is rational.

References

Airiau, S., Bonzon, E., Endriss, U., Maudet, N., & Rossit, J. (2016). Rationalisation of Profiles of Abstract Argumentation Frameworks. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS)* (pp. 350–357).

Bench-Capon, T. (2003). Persuasion in Practical Argument Using Value-Based Argumentation Frameworks. *Journal of Logic and Computation*, *13*(3), 429–448.

Dung, P. M. (1995). On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, *77*(2), 321–357.

Fan, X., & Toni, F. (2015). On Explanations for Non-Acceptable Arguments. In *International Workshop on Theory and Applications of Formal Argumentation* (pp. 112–127).

Conservativeness, Language, and Deflationary Metaontology

Jonas Raab

University of Manchester, Manchester, UK
jonas_raab@web.de

Abstract. This paper is about the role conservativeness plays in the deflationary metaontologies of Schiffer and Thomasson. Deflationary metaontologies lead to trivial answers to existence questions. However, to do so, they rely on languages and their possible extensions. To rule out inconsistent (or otherwise damaging) extensions, the common move is to restrict extensions to conservative ones. I argue that it is this very move that leads to trouble for both Schiffer's and Thomassons's account.

Keywords: Conservativeness · Deflationary Metaontology · Easy Ontology · Amie L. Thomasson · Pleonastic Entities · Stephen Schiffer.

1 Introduction

Ever since Field's 1980 seminal *Science Without Numbers*, the idea of using *conservativeness* in ontological debates has been present. Field's strategy for nominalism with respect to mathematical entities is to present an appropriate base theory and to show that adding the mathematical framework leads to a conservative extension of the base theory.

To get a better understanding of this, we need to know what 'conservative' means; we can (roughly) define it as follows: Let T and T' be theories. Let $\mathcal{L}_T$ be the language of T understood as a set of well-formed formulas. Then:

(CONS) T' is a *conservative extension* of T if, and only if, (iff)
for every $S \in \mathcal{L}_T$, if S follows from T', then S follows from T.

For our purposes, it does not matter how exactly to understand 'follows from'. Note, however, that there are several possibilities and they come apart (see Shapiro's [14], and Field's [6] and [8, pp. P-16ff.]). As the definition implies, the language $\mathcal{L}_{T'}$ of T' is an extension of $\mathcal{L}_T$, i.e., $\mathcal{L}_T \subseteq \mathcal{L}_{T'}$. In particular, every theory is a conservative extension of itself and every extension of an inconsistent theory is conservative.

Field's idea behind invoking conservativeness is to show that, in contrast to "theoretical entities in physics" [8, p. 7], "mathematical entities are not theoretically indispensable" [8, p. 8]. If we have an appropriate theory that does not invoke mathematical entities, we can see that we don't need such entities. However, having mathematics at our disposal makes proving easier; it is dispensable,

but still useful [8, p. 16]. In particular, Field is not arguing that we should not invoke mathematics when doing physics (and other sciences); but if his program is successful, we can see that mathematics does not need to be *true* to do its job [8, p. 8]—it only needs to make things easier.

Conservativeness resurfaces in the neo-Fregean debate. Frege's original attempt to derive arithmetic from (second-order) logic and definitions failed because of his infamous *Basic Law V*. Russell showed that it implies a contradiction, i.e., that on its basis what's now known as *Russell's Paradox* is derivable. Basic Law V is a so-called abstraction principle which has the form of a biconditional whose one side is an identity statement between singular terms which are introduced via a term-building operation (the abstraction operator, hence the name 'abstraction principle') and whose other side involves an equivalence-relation:

$$@xFx = @yGy \leftrightarrow F \equiv G \tag{AP}$$

where '$@x$' is the abstraction operator and '$\equiv$' is an equivalence relation (see Cook [2] for a more general characterization of abstraction principles).

Neo-Fregeans try to do better than Frege. In particular, as the work of Wright [17] and Boolos [1] show, Frege's program is consistent if one invokes what is known as *Hume's Principle* instead of Basic Law V. Frege himself only uses the latter to derive the former. Hume's Principle is itself an abstraction principle and the neo-Fregeans base their entire ontology on such principles. The idea behind abstraction principles is that they do not create any new commitments, but only 're-carve' what's already there; in this manner, we just get new concepts [9, §64]. But, as it turns out, different abstraction principles can't be true together [4, p. 21]. This is the well-known bad company problem [10]. One thread of neo-Fregean philosophy consists in distinguishing the good from the bad abstraction principles; one suggestion to accomplish this is to only allow for abstraction principles that are conservative [10, p. 325].

More recently, conservativeness has found a place in deflationary metaontologies. Both Schiffer [13] and Thomasson [16] invoke the concept. And it is their use that is the focus of this paper.

In the following, I consider Schiffer's and Thomasson's accounts. I first give a brief overview of how they want to derive entities (Section 2) and give their respective motivations for invoking *conservativeness*. Then, I criticize Schiffer's and Thomasson's accounts in Sections 2.1 and 2.2, respectively. I conclude this paper (Section 3) by briefly considering whether my criticisms also affect Field's program and the neo-Fregean account.

2 Deflationary Metaontology and Conservativeness

Deflationary metaontologies proceed by looking at how language lets us infer disputed entities; these inferences are supposedly 'easy'. Thomasson characterizes any approach as 'easy' if it has "two features" [16, p. 128]:

(1) "all well-formed existence questions may be answered by conceptual and/or empirical work" [16, p. 128], and

(2) "at least some disputed existence questions may be answered by means of trivial inferences from uncontroversial premises" [16, p. 128].

Thus, she characterizes both the neo-Fregean approach as well as Schiffer's as 'easy' [16, p. 127]; the former uses established truths to introduce new concepts via abstraction, the latter invokes trivial inferences to derive 'pleonastic entities' which are "entities whose existence is secured by something-from-nothing transformations" [13, p. 51]. Schiffer also explains why he uses the term 'pleonastic entities': "something-from-nothing transformations often take us to pleonastic equivalents of the statements from which they are inferred" [13, p. 51]. The idea is that the inferences just redundantly restate what the premises already contain.

Similarly, Thomasson relies on easy arguments to derive certain entities. For example, she derives the existence of properties as follows:

> From a sentence like 'the table is brown' we may also infer 'the table has the property of brownness', and thus that there is a property[.] [16, pp. 102f.]

In similar fashion, Schiffer explains:

> Consider the property of being a dog. On my view, there isn't a lot more to this property than can be culled from the something-from-nothing transformation that allows us to move back and forth between
> (3) x is a dog
> and its pleonastic equivalent
> (4) x has the property of being a dog. [13, pp. 64f., numbering changed]

We can note two things here. Firstly, Schiffer has to invoke "certain qualifications" [13, p. 72] to his approach to not run into Russell's Paradox (see also [12, pp. 164ff.]). For, in similar fashion to the above something-from-nothing transformations from (3) to (4) and vice versa, we can arrive at the equivalence of 'the property of being a dog is not self-instantiating' (since the property is not a dog) and 'the property of being a dog has the property of not being self-instantiating'. Therefore, we arrive at the property of not being self-instantiating—a property that gives rise to the paradox.

Thomasson's easy arguments are threatened by similar considerations. Nevertheless, for the purposes of this paper, I waive this problem.

Secondly, easy arguments bring with them the threat of over-generating entities, i.e., a bad company problem. Schiffer gives the following example:

> x is a *wishdate* $=_{df}$ x is a person whose existence supervenes on someone's wishing for a date, every such wish bringing into existence a person to date. [13, p. 53, his emphasis]

This is something that we don't want in our ontology. However, in some sense this is harmless, viz., it does not lead to an outright contradiction. It is Eklund [5, pp. 100ff.] who introduces conditions for entities that would lead to contradiction if they all existed. Thus, there is the threat of contradiction after all.

Thomasson [16, ch. 8], too, discusses whether there is a 'bad company' problem for her approach. Both she and Schiffer invoke conservativeness as a way to rescue their approaches. Schiffer draws a distinction between the good and the bad cases by noting that the former are conservative whereas the latter are not:

> There are numerous theories T such that when we add to T the concept of a wishdate together with the claim that wishing for a date entails the existence of a wishdate, the resulting theory is not a conservative extension of T. [13, p. 54]

Thomasson, on the other hand, puts a conservativeness restriction on the introduction of new terms to a given language L:

> Introducing [a] term must not analytically entail anything statable in unextended L that was not already analytically entailed by truths stated in L. (This is a version of the familiar conservativeness requirement [...]). [16, pp. 263f.]

So, both Thomasson and Schiffer put a restriction on what counts as a proper extension to block the bad company problem; as the bad terms lead to extensions that are not conservative, they are dismissed.

In the following, we take a closer look at how Schiffer and Thomasson want to use conservativeness. We start with Schiffer's account.

2.1 Schiffer and Conservativeness

In introducing conservativeness, Schiffer notes that there are different reasons why certain extensions of a theory are not conservative [13, p. 55]. For example, the extended theory "may entail that more than such-and-such many things exist" [13, p. 55]; but such numerical claims are already statable in the base theory, so that the extension is not conservative because it contradicts the count of the base theory. Field, on the other hand, notes that adding some platonistic theory to a nominalistic one can also lead to contradiction, namely in those cases in which the nominalistic theory "may say things that rule out the existence of abstract entities" [8, p. 10] (see also [13, p. 55]). Schiffer [13, pp. 56f.] follows Field's strategy to circumvent this problem, viz., appropriately restricting the quantifiers of the base theory:

> For a theory or sentence T, $T^{\neg F}$ is the theory or sentence that results from restricting each quantifier in T to things that aren't F. [13, p. 57][1]

He puts this idea to use in the following way:

> I now offer the following conservative-extension criterion for being a pleonastic concept.

[1] Note that I changed the notation: Schiffer uses '$\sim F$' where I am using '$\neg F$'. Field's [8] notation for the restriction of a theory 'T' is 'T^*'.

> **(CE)** The concept of an F implies true something-from-nothing F-entailment claims—and is therefore a *pleonastic concept*—iff (i) it implies something-from-nothing F-entailment claims, and (ii) for any theory T and sentence S expressible in T, if the theory obtained by adding to $T^{\neg F}$ the concept of an F, together with its something-from-nothing F-entailment claims, logically entails $S^{\neg F}$, then $T^{\neg F}$ logically entails $S^{\neg F}$.
>
> In other words, adding pleonastic entities to any theory conservatively extends that theory, relative to the restriction on quantification. [13, p. 57, his emphasis]

Thus, having the restriction in place, the extension does not imply any contradictions. The old part is appropriately characterized by the qualification of '$\neg F$', but we can still drop the restriction and talk about all the things in the extended theory (which will be symbolized as '$T^{\oplus F}$'). To fully appreciate the definition, we need also to know what 'something-from-nothing F-entailment claims' are:

> Where '$\Rightarrow$' expresses metaphysical entailment, '$S \Rightarrow \exists xFx$' is a *something-from-nothing F-entailment claim* iff (i) its antecedent is metaphysically possible but doesn't *logically* entail either its consequent or any statement of the form '$\exists x(x = \alpha)$', where 'α' refers to an F, and (ii) the concept of an F is such that if there are Fs, then $S \Rightarrow \exists xFx$. (I'll say that the concept of an F 'implies' a something-from-nothing F-entailment claim if it satisfies (ii).) [13, pp. 56f., his emphases]

Schiffer explains the metaphysical entailment as follows: "*A metaphysically entails B* just in case the material conditional $A \rightarrow B$ is metaphysically necessary" [13, p. 56, n. 7, his emphasis]. So, the idea is that the introduction of the pleonastic concept F guarantees the existence of Fs and nothing else. As the existence of an F is clearly not expressible in the base theory (it lacks the predicate 'F'), (**CE**) is satisfied, i.e., the extension is conservative.

However, we can problematize this right away. As the quotation states, condition (ii) is a *conditional* whose antecedent involves the *existence of Fs*. Thus, if there are no Fs, then condition (ii) is vacuously satisfied. If condition (i) is satisfied then, according to the definition, '$S \Rightarrow \exists xFx$' is a something-from-nothing F-entailment. But, as there are no Fs, the vacuous condition (ii) does not guarantee the truth of '$\exists xFx$'. This means that we have no reason to assume non-empty something-from-nothing F-entailment claims *unless* we have independent reasons to believe that there are Fs, i.e., that the antecedent of condition (ii) is satisfied. Therefore, assuming '$S \Rightarrow \exists xFx$' to be a *non-empty* something-from-nothing F-entailment claim begs the question as it *presupposes* the existence of Fs.

Note that, if there are no Fs, the 'S' in '$S \Rightarrow \exists xFx$' is a sentence of $\mathcal{L}_T$ as otherwise there is the danger that condition (i) is not satisfied, i.e., that it implies the existence of Fs. Assuming T to decide the sentence, we can see that it *cannot* be a consequence of T as it would be a consequence of $T^{\oplus F}$ so that it follows from $T^{\oplus F}$ that $\exists xFx$—even if there are no Fs.

Further, as (**CE**) demands, a pleonastic concept F must entail something-from-nothing F-entailment claims and give rise to a *conservative* extension of $T^{\neg F}$. Then, if there are no Fs, '$S \Rightarrow \exists x Fx$' is an empty something-from-nothing F-entailment. Further, as this 'metaphysical entailment' does not entail the existence of Fs, the extension *must* be conservative as there is *no* logical consequence that is not already one of $T^{\neg F}$; among the *new* consequences would be '$\exists x Fx$' which, by assumption, is not the case. So, if we want to actually introduce Fs, we have to assume the existence of Fs; but such an assumption is clearly question begging. In terms of the theories, this means that the base theory already contains Fs. But if this is the case, then there are cases in which $T^{\neg F}$ is *inconsistent* as the restriction of a sentence of the form '$\exists x(\ldots Fx \ldots)$' is '$\exists x(\neg Fx \wedge \ldots Fx \ldots)$', i.e., $T^{\neg F}$ demands there to be Fs that are not Fs. Thus, the strategy to make the conservativeness requirement work stands in the way of successfully introducing Fs to T.

Before getting to the next criticism, let me point out two things. Firstly, the talk of 'material conditional' in the quotation above as well as the qualifications Schiffer invokes to not run into Russell's Paradox [13, p. 72] seem to imply that he is working with *classical logic*.[2] His treatment of vagueness [13, ch. 5] puts that into doubt, though. Thus, we need to consider both cases.

Secondly and because of the above, we need a better understanding of what exactly is going on when a theory is extended. A theory T is a set of sentences in the language $\mathcal{L}_T$ that is closed under logical consequence '$\vdash_L$' where the subscript 'L' indicates the logic, i.e., if A_T is the base of T (e.g., an axiomatization of T), then $T = \{S \in \mathcal{L}_T | A_T \vdash_L S\}$. The restricted theory $T^{\neg F}$, however, does not put a restriction on the *logical theorems*, i.e., on sentences S such that $\emptyset \vdash_L S$ since we do not restrict the logic. That means that we can distinguish between the logical laws and other sentences in $T^{\neg F}$ via checking whether or not the sentence involves a quantifier restricted to '$\neg F$'.

With this background, we can ask the crucial question what these newly introduced/inferred entities are like, and, in particular, whether these entities are, indeed, *new*. Schiffer claims the following:

> Thus, as with all pleonastic entities, properties have 'no hidden and substantial nature for a theory to uncover.' The essential truths about them are directly or indirectly determined by the hypostatizing practices constitutive of the concept of a property, together with those necessary a priori truths applicable to things of any kind, such as that if $x = y$, then whatever property x has, y has, and vice versa. As regards the principles by which properties are individuated, it means that if a question of individuation is left unsettled by the practices constitutive of the concept of a property, then that question has no determinate answer. [13, p. 63]

Apparently, all we need to know are 'constitutive practices' to arrive at the properties of these new things in our (now extended) theory. However, the point that

[2] One way to circumvent the threat of, e.g., Russell's Paradox is to *weaken* classical logic; e.g., Field [7] opts for a non-classical logic to "save" the so-called *Truth Schema*.

I want to problematize is the application of 'necessary a priori truths applicable to things of any kind'. Schiffer relies on this when he says that "Leibniz's law gives us a means for establishing numerous non-identities" [13, p. 63]. The problem is simply this: (i) either Leibniz's law (as well as all other necessary a priori truths) is just another sentence of the base theory, or (ii) it is a logical theorem; both options are problematic, though. Let us consider these options in turn.

Suppose that (i) is the case. Then Leibniz's law, i.e. the sentence(-schema) '$\forall x \forall y(x = y \to (\varphi(x) \to \varphi(y)))$' is part of the base theory T. The example Schiffer uses himself is the something-from-nothing transformation from 'Jane was born on a Tuesday' to 'Jane's birth was on a Tuesday' [13, p. 63]. The question he asks is whether or not Jane's birth is the same as her death. To introduce the term 'birth' ('B') to T, we have to restrict the quantifiers in T to '$\neg B$', i.e., we move to the theory $T^{\neg B}$. Thus, Leibniz's law in $T^{\neg B}$ is '$\forall x \forall y(\neg B(x) \wedge \neg B(y) \to (x = y \to (\varphi(x) \to \varphi(y))))$', i.e., it only applies to entities that are *not* births; and this does not change when we add the 'B' to $T^{\neg B}$. This, however, stands now in the way of comparing Jane's death with her birth. The latter is not in the range of the restricted quantifiers and, therefore, Schiffer's argument for the distinctness of the two fails. The quantifiers ('$\forall x$', '$\forall y$') still range over the whole domain and so range over births, but the law itself is not applicable in answering identity questions regarding the new term.

Thus, to make Schiffer's application of, for example, Leibniz's law work, we have to opt for option (ii), i.e., we understand Leibniz's law as well as all the 'necessary a priori truths' to be *logical theorems*.

As Schiffer does not provide a list of the necessary a priori truths, we do not know exactly what principles he counts as such. However, he says the following:

> What generates the conflict in the first place is the status that [...] the law of excluded middle [...] [has] in our conceptual repertoire. The underived conceptual [role] of our [notion] of [...] disjunction dispose[s] us to accept instances of excluded middle[.] [13, p. 225]

This suggests that the following law of excluded middle (LEM) is a conceptual truth:

$$\varphi \vee \neg\varphi. \tag{LEM}$$

Now, it is clear that Schiffer *cannot* take (LEM) to be a necessary a priori truth. For, as he claims, a question regarding 'individuation' that is "left unsettled by the practice constitutive of the concept of a property [...] has *no determinate answer*" [13, p. 63, my emphasis]. But this would be simply false if (LEM) was true; for, every answer would be determinate.

In his discussion of vagueness, Schiffer [13, p. 227] explicitly rejects (LEM). Thus, he rejects classical logic. Again, as he speaks in one of the quotations above of 'material conditional', it is rather unclear how we are to understand the 'something-from-nothing F-entailment claims' as well as (**CE**) then.

But let us waive this problem here, i.e., assume that Schiffer is endorsing a *non-classical logic*, and consider the application of Leibniz's law again. He argued that because "Jane's birth occurred in 1850 and her death occurred in 1933" [13,

p. 63], the birth and the death must be non-identical. But, as (LEM) does not hold, just because Jane's birth occurred in 1850 does not imply that Jane's birth did not occur at any other time; it might have occurred at several different times. Similarly in the case of her death. Thus, we need *additional* information such as 'Jane's birth did *only* occur in 1850' to apply Leibniz's law. Thus, without this additional information, it is not given that these entities are distinct. But if they aren't, we have not successfully introduces anything *new*; the conservativeness requirement stands in the way once more.[3]

2.2 Thomasson and Conservativeness

This brings us to Thomasson's account. She does not discuss the above inconsistency issues related to extensions; let us assume that we do not run into such, i.e., that the extensions are conservative. Still, she runs into similar problems as Schiffer above as will become clear in the following. Let me first give a rough outline of her approach before criticizing it and explaining the problem.

The central idea for Thomasson's easy approach is captured in the following deflationary principle:

(E) "Ks exist iff the application conditions actually associated with 'K' are fulfilled." [16, p. 86]

For our purposes, the 'actually associated' bit is not important (see [16, pp. 85f.] for the motivation). What is important are the application conditions. We do not need to give a full account, but only note the underlying motivation.

Already in her [15], Thomasson introduces application conditions to circumvent what is known as the *qua* problem [3, pp. 79ff.] for (purely) causal theories of reference. In a nutshell, the problem is that unless we have "some very basic concept of what sort of thing" [15, p. 38] we are referring to, reference is indeterminate. Her solution to this problem is that "nominative terms must be associated with a sortal or, more generally, categorial concept" [15, p. 39]; she calls this a 'hybrid' account [15, p. 48]. All this is confirmed in the more recent [16, p. 95].

However, not every concept categorizes as categorial: terms such as 'individual', 'object', or 'thing' are not among them [15, p. 42]. This means that they cannot be invoked in application conditions.

So, to guarantee successful reference using a term 'K', we have to have application conditions associated with 'K' to ground the reference. If these are fulfilled, (**E**) guarantees the existence of the corresponding entities.

We can raise the following problem here. To have grounded reference, we already need successful reference grounding. For, any term that we want to

[3] Note [13, pp. 233ff.] where Schiffer discusses arguments that attempt to show that Leibniz's law is violated. As his endorsement of a non-classical logic is motivated by his treatment of vagueness, and he says that "almost every expression is to some extent vague" [13, p. 178], his own discussion is highly relevant for his former application of Leibniz's law. The problem is that his own discussion stands in the way of such a simple application.

introduce is in need of application conditions to disambiguate the reference. But the application condition must involve a categorial term to successfully disambiguate. Now, how did we get this categorial term? Apparently, by means of *yet another* categorial term, and so on.

Assuming that we don't allow to go full circle (as Thomasson rejects with her condition (4) in [16, p. 96]), we can only stop this regress by assuming that we started with a set of *referring*[4] categorial terms that have not been given to us by application conditions.

Moreover, given the need to introduce new terms via associating appropriate application conditions to them, we can also note that Thomasson cannot start from a language not containing any predicates (i.e., terms). For, there is just nothing in there to disambiguate the reference of a new term. Note, however, that, trivially, adding terms to an empty language is conservative.

Now, this leads to a problem for Thomasson's account of application conditions. Principle (**E**) is supposed to apply *across the board*; recall (1) from above. We want to answer *all* well-formed existence questions. To be a well-formed existence question, the terms involved must have application conditions associated with them [16, p. 219]. But this means that even the terms in the set of terms with which we have to start before extending them with new terms must have application conditions. Presumably, we can ask and answer existence questions regarding them even in the unextended base language. Thus, their application conditions must be statable in terms of one another; otherwise (1) is not satisfied. Thomasson also makes this a condition for introducing new terms to an "unextended language L" [16, p. 263]:

> The term(s) must be introduced via a conditional that gives sufficient conditions for its(/their) application, stated *using the extant terms of L and/or other minimally introduced terms*. [16, p. 263, my emphasis]

But this means that we must have terms 'K' and 'K'' whose application conditions involve one another.[5] So, to answer the question whether there are Ks, we have to check whether there are K's; and to check whether there are K's, we have to see whether there are Ks.

Indeed, Thomasson distinguishes 'basic' from 'derivative' terms:

> Once basic nouns are in place [...] we can introduce new nouns *on the basis of others*. [16, p. 99, my emphasis]

The above circularity problem concerns the basic terms. I waive this problem for the moment.[6] There is, however, another problem, viz., whether we can even

[4] If they were non-referring, we could never establish reference. For, we use the categorial term to specify what something is, and then introduce further conditions to be more specific. But, if the categorial term did not refer, neither would the more specific one.

[5] Strictly speaking, it means that for the basic terms 'K_0', 'K_1', ..., 'K_n' we end up in a similar situation. For simplicity, I chose and claimed this about two.

[6] I discuss it in detail in [11].

introduce all the other terms that correspond to entities that Thomasson wants to infer. Thomasson characterizes the basic terms as those that "we tend to learn early in our cognitive and linguistic development" [16, p. 104]; as examples, Thomasson refers to "what Carnap called 'the thing language' [...] such as 'piece of paper', 'desk', and the like" [16, p. 106]. Her examples do not include any abstracta—which seems reasonable given her characterization of 'basic'.

So, the question to be asked now is how Thomasson is able to infer the existence of abstracta such as numbers. We can note that Thomasson does not think of 'number' as a basic term [16, p. 217]. However, she does think that she is inferring abstracta such as numbers and properties (for the latter, see [16, pp. 102f.] quoted above in Section 2).

To answer the existence question regarding numbers, we have to look at the application conditions of 'number'. Since we just saw that this is a derivative term, it must have been introduced by means of other terms. We can also note that numbers are abstract, so that either (i) a term that is involved in the application conditions of 'number' is associated with an abstractum, or (ii) the application conditions involve a phrase like '... and is abstract' or 'abstract ...'. Let us look at the options in turn.[7]

Suppose first that (i) is the case. Then we can apply the same reasoning again. Opting always for this option, we must bottom out in basic terms. Thus, there must be a term among the basic terms that is associated with an abstractum. This is so because of the conservativeness requirement together with Thomasson's requirement to have application conditions in place. However, this is rather implausible given the characterization of 'basic'.

Here we can see that Thomasson does not rely on having the concept '$\neg K$' to introduce 'K'. However, this is part of what creates the problem here. Thus, we end up with the following dilemma: to even be in a place to have a conservative extension, we need to invoke concepts from outside the theory we want to extend (given the characterization of 'basic' terms that constitute the base theory and the condition quoted above on using the vocabulary of the base theory), or we are not in a place to extend the theory to all the concepts we want, as will become clear in the following.

So let us take a look at (ii). Again, because of the conservativeness requirement, we cannot be sure whether or not the application conditions are fulfilled now. Suppose, for example, that the application conditions of 'number' are 'abstract K', where 'K' is a (combination of) basic term(s) (i.e., associated with something non-abstract). How shall we check now whether or not these application conditions are fulfilled?[8] Just assuming them to be fulfilled is obviously question-begging—as it already was in Schiffer's account.

[7] Note that the requirement quoted above means that we introduce conditionals. Nonetheless, I speak of application conditions to mean the antecedent of such.

[8] Note, too, that using Hume's principle as (part of) the application conditions for 'number' does not help here. Hume's principle refers to *relations*, so that we can ask the question about how we got the term 'relation', and the game repeats itself as we are back in case (i).

Summing up, the circularity issue puts Thomasson's condition (1) into question as her own approach is not applicable to *all* existence questions. She might rebut, however, that the existence questions in question here are not *well-formed*. The problem with this, though, is that this infects all other existence questions that have to rely on terms that don't have application conditions associated with them but nonetheless correspond to entities (see [11] for details).

The abstractness problem, on the other hand, might conflict with (2). Thomasson says that "at least some" [16, p. 128] questions have to be answered in this way, but as my above argument shows, none of these "disputed entities" [16, p. 128]—if abstract—can be derived if we don't already start with abstracta. In particular, by the conservativeness requirement that she invokes together with her characterization of 'basic' terms, Easy Ontology as developed by her is not capable of inferring *any* abstract entities. As concrete entities are usually not disputed (or at least not in the same way), Easy Ontology does not fulfil (2).[9]

3 Conclusion

To sum up, both Schiffer and Thomasson invoke conservativeness in their deflationary metaontologies. However, both end up with problems that arise therefrom. Even if we successfully introduce new terms to our language, it is rather unclear what properties the corresponding entities have. In particular, it is consistent to assume them to have properties that are not normally ascribed to them. The reason for this is simply that we have to restrict the quantifiers to introduce something new; but it is then an open question what exactly has been introduced. Just assuming that it has certain properties is question begging in an ontological context.

Further, it is rather unclear whether we can *infer* the existence of certain disputed entities. As the discussion of Thomasson's work makes clear, to even get to abstracta, we have to assume (appropriately many) abstracta to introduce others. However, this is not the result that her easy approach promises.

We can also observe that Field's approach is not in any way at risk to similar worries. The reason is simply that his project does not try to introduce new terms/entities, but he looks at the full picture and singles out a base that is appropriate for the job at hand. We can describe this as top-down, and nothing has been said that dismisses a successful top-down approach. The bottom-up approach as exemplified by the discussed deflationary metaontologies faces the difficulties pointed out.

We can also note that the restriction to the negation of the concept that we want to introduce is not a problem for the top-down approach; the concept was already there to be invoked. However, the bottom-up approach seems to force us to already have the very concept we want to introduce.

Lastly, the neo-Fregean approach understood as a 're-carving' of what's already there might also be unaffected by the problems the deflationary meta-

[9] In [11], I argue this point more fully to conclude that even empirical inquiry is impossible by the Easy Ontologist's *own lights*.

ontologies of Schiffer and Thomasson face. If we understand everything to be already there, the new concepts are introduced via biconditionals that allow us to understand the *new* concept by recourse to the old ones. And as abstracta are already assumed, neo-Fregeans do not run into the problem of not being able to introduce any abstract entities. The problem facing Schiffer's account is also not a threat to them. The reason is again that we are not extending any theory (and with that, any domain), and the logic is always *classical*.

Acknowledgments

I would like to thank Chris Daly, Andries de Jong, and Jeroen Smid for helpful comments and discussion.

References

1. Boolos, G.: Saving Frege from Contradiction. Proceedings of the Aristotelian Society **87**, 137–151 (1986/87)
2. Cook, R. T.: Conservativeness, Cardinality, and Bad Company. In: Ebert, P. A. and Rossberg, M. (eds.): Abstractionism: Essays in the Philosophy of Mathematics, pp. 223–246. Oxford University Press (2016)
3. Devitt, M., Sterelny, K.: Language and Reality. An Introduction to the Philosophy of Language. 2nd edn. The MIT Press (1999)
4. Ebert, P. A., Rossberg, M.: Introduction to Abstractionism. In: Ebert, P. A. and Rossberg, M. (eds.): Abstractionism: Essays in the Philosophy of Mathematics pp. 3–33. Oxford University Press (2016)
5. Eklund, M.: Neo-Fregean Ontology. Philosophical Perspectives **20**(1), 95–121 (2006)
6. Field, H.: On Conservativeness and Incompleteness. In: Field, H.: Realism, Mathematics and Modality, pp. 125–146. Basil Blackwell, Oxford (1989). Originally published in Journal of Philosophy **81**(5), 239–260 (1985)
7. Field, H.: Saving the Truth Schema from Paradox. Journal of Philosophical Logic **31**(1), 1–27 (2002)
8. Field, H.: Science Without Numbers: A Defense of Nominalism. 2nd edn. Oxford University Press (2016)
9. Frege, G.: Die Grundlagen der Arithmetik. Eine logische Untersuchung über den Begriff der Zahl. Mit einem Nachwort herausgegeben von J. Schulte. Reclam, Stuttgart (1989)
10. Linnebo, Ø.: Introduction. Synthese **170**(3), 321–329 (2009)
11. Raab, J.: The Unbearable Circularity of Easy Ontology. Manuscript. Available at: https://www.academia.edu/36916145/The_Unbearable_Circularity_of_Easy_Ontology
12. Schiffer, S.: Language-Created Language-Independent Entities. Philosophical Topics **24**(1), 149–167 (1996)
13. Schiffer, S.: The Things We Mean. Oxford University Press (2003)
14. Shapiro, S.: Conservativeness and Incompleteness. The Journal of Philosophy **80**(9), 521–531 (1983)
15. Thomasson, A. L.: Ordinary Objects. Oxford University Press (2007)
16. Thomasson, A. L.: Ontology Made Easy. Oxford University Press (2015)
17. Wright, C.: Frege's Conception of Numbers as Objects. Aberdeen University Press (1983)

Interpreting Intensifiers for Relative Adjectives: Comparing Models and Theories

Zhuoye Zhao

ILLC, University of Amsterdam

Abstract. Adjectives such as *tall* or *late* which can enter comparative constructions or be modified by intensifiers such as *very* are called *gradable*. They have received considerable attention in formal semantics and, more recently, in Bayesian pragmatics. While comparative constructions are well understood, less is known about the contribution of intensifiers. In this paper, we compare several concrete models for the meaning of *very tall* with data from a previous study.

1 Introduction

1.1 Degree Semantics

Degree Semantics (Kennedy & McNally, 2005; Kennedy, 2007, among others) proposes that gradable adjectives such as *tall* and *late* map individuals to *degrees* on a *scale*. For example, $[\![\text{tall}]\!] = \lambda x.\mathbf{height}(x)$, which gives a map from individuals to their degree of *tallness*. When used in positive forms as in *Ronald is tall*, which is obtained by combining *tall* with a null morpheme *pos*, it means the *degree* of *tallness* (i.e. height) exceeds a context-dependent threshold $\theta(C)$.

(1) $[\![pos\ \text{tall}]\!] = \lambda x.\mathbf{height}(x) \geq \theta(C)$

The threshold $\theta(C)$ is determined contextually in the sense that it has different values with respect to different *comparison classes* (encoded by the argument C). For instance, *a tall tree* and *a tall man* are definitely judged by different standard. The Bayesian models we will introduce shortly will show ways in which the threshold is selected according to C.

Different adjectives differ in the type of measure functions they denote and the associated scale structure. Kennedy & McNally (2005) distinguish between *absolute adjectives* like *late* or *full*, which map individuals onto a degree scale that is lower- or upper-bound, and *relative adjectives* like *tall* which have an open scale. Scale structure affects $\theta(C)$: absolute adjectives tend to pick the scale endpoint as their threshold, which is known and does not depend much on the comparison class. Open scales do not provide such a salient threshold, so their $\theta(C)$ is more context-dependent and less certain. As a consequence, relative adjectives tend to be *vague*. For example, it is hard to decide whether a 5ft10in tall man is tall. By contrast, adjectives like *late* or *full* show no such vagueness – someone is *late* as long as they show up after the scheduled time.

Having introduced a comparison threshold $\theta(C)$ into the semantics of gradable adjectives, the meaning of degree modifiers such as *very* follows naturally – they shift the threshold $\theta(C)$ to a higher value. Klein (1980) proposed a formal semantic account for degree modifiers that captures the threshold shifting and can be easily adapted into the notions of degree semantics. The basic idea is that a sentence like *Ronald is very tall* is true if Ronald is tall compared to the set of tall people. We can formally define it as follows:

(2) $[\![\text{very } pos \text{ tall}]\!] = \lambda x.\mathbf{height}(x) \geq \theta(C')$,
where $C' = \{x \mid [\![pos \text{ tall}]\!](x) = 1\}$

One of the goals of this paper is to test the prediction of this account using probabilistic pragmatic models (see §1.2), and compare it with another purely pragmatic account proposed by Bennett & Goodman (2015) (also see §1.2).

1.2 Probabilistic Pragmatic Models for Gradable Adjectives

In order to fully capture the meaning of gradable adjectives, a mechanism to determine or infer the context-dependent threshold $\theta(C)$ is necessary. Fortunately, probabilistic pragmatic models (in the sense of Franke & Jäger, 2016) have been successful in making quantitative predictions for $\theta(C)$, and in further explaining the linguistic phenomena we're interested in. The basic assumption of such models is that language users are goal-oriented Bayesian agents that are involved in social interactions where speaker and listener communicate and recursively reason about each others' goals/inferences. Each utterance u is attributed probability $P(u|w)$ to be chosen by a speaker with knowledge-state w under the assumption that the speaker is trying to maximize some notion of utility (the definition of which varies in different models). Listeners simply inverse the probability using Bayes' rule to obtain a probability distribution on possible states of the world given what the speaker said: $P(w|u)$.

Here we present two probabilistic pragmatic models for gradable adjectives, the Rational Speech-Act Model (RSA) proposed by Lassiter & Goodman (2014), and the Speaker-Oriented Model (SOM) by Qing & Franke (2014). Specifically, the RSA model is a *listener-oriented* model, which predicts the threshold $\theta(C)$ as an inference of a (pragmatic) listener, whereas the SOM model, as indicated by the name, derives $\theta(C)$ at the speaker's level based on his/her prior knowledge about the world. We will now present the main features of these two models. In the following we focus on the relative adjective *tall* and the utterance '*Ronald is tall*', marked as u and the trivial empty utterance u_0.

Rational Speech-Act model (RSA)

The essential idea behind (strongly) Bayesian models of pragmatics is that the listener uses Bayes' rule to recover a speaker's knowledge state w in a context C given the speaker's utterance u.

(3) $P(w|u, C) \propto P(u|w, C) \times P(w)$

In our case, we are only interested in Ronald's height h_0, so w can be reduced to the speaker's knowledge of this height. We assume that the listener has a prior knowledge of the distribution of possible heights, $\phi(h)$. The only missing part now is a model of how the speaker choose their utterance, i.e. $P(u|w, C)$.

The idea behind the RSA is that the listener models a speaker who tries to minimize the cost of their utterance, while maximizing informativity for a virtual "literal listener" L_0, who simply updates the prior by conditioning on u being literally true. This is where the semantics come into play: Lassiter & Goodman (2014) assume the standard Degree semantics, so the truth of u depends on a threshold θ. In their model, the listener has no prior knowledge of θ, but assumes that the speaker has exact knowledge of it. θ is thus a free parameter that the listener must infer together with h_0, Ronald's actual height.

Concretely, informativity is defined as negative surprisal value of L_0's belief about h_0 after hearing u, and the speaker's utility function is as (4):

(4) $U_{\text{rsa}}(u, \theta, h_0) = \log(\phi(h_0|u, \theta)) - \text{Cost}(u)$

The speaker tries to maximize utility, but is assumed to do it in a sub-optimal fashion (using a soft-max with parameter $\lambda < \infty$):

(5) $\sigma(u|\theta, h_0) \propto exp(\lambda \cdot U_{\text{rsa}}(u, \theta, h_0))$

The listener then infers a joint distribution for θ and h by applying Bayes' rules. Here we will only be interested in the posterior distribution of θ, which is given by the formula in (6).

(6) With $c_{\text{rsa}} = \text{Cost}(u)$ and Pr the (uninformative) prior on θ:

$$\rho(\theta|u) \propto \int_{-\infty}^{\infty} \phi(h) \cdot Pr(\theta) \cdot \sigma(u|h, \theta) dh = \frac{Pr(\theta) \cdot \int_{\theta}^{\infty} \phi(h) dh}{1 + e^{\lambda c_{\text{rsa}}} \cdot (\int_{\theta}^{\infty} \phi(h) dh)^{\lambda}}$$

Speaker-Oriented Model (SOM)

The SOM differs from the RSA model in that instead of the literal listener L_0, it assumes a listener L sharing the prior knowledge $\phi(h)$ with the speaker. Moreover, instead of a fixed value of the threshold $\theta(C)$, it provides a mechanism to derive the probability of the speaker using a specific θ, hence gives a generalization over possible contexts.

Concretely, keeping the assumption that the speaker tries to maximize the utility, the SOM replaces *informativity* with the notion of *Expected Success* (given by the expected value of the probability of L successfully guessing the actual height h_0), and replaces the cost function with its marginalization over all possible heights $h > \theta$.

(7) With the cost parameter c,

$$\begin{aligned} U_{\text{som}}(\theta) &= ES(\theta) - \text{Cost}(u) \\ &= \int_{-\infty}^{\theta} \phi(h)\phi(h|u_0, \theta) dh + \int_{\theta}^{\infty} \phi(h)\phi(h|u, \theta) dh - \int_{\theta}^{\infty} \phi(h) \cdot c_{\text{som}} dh \end{aligned}$$

Again, the threshold θ is chosen sub-optimally as in (5):

(8) $Pr(\theta) \propto exp(\lambda \cdot U(\theta))$

Then according to Degree semantics, the probability of using the utterance u can be given as the probability of $\theta \leq h_0$:

(9) $\sigma(u|h_0) = P(\theta \leq h_0) = \int_{-\infty}^{h_0} Pr(\theta) d\theta$

1.3 Bennett & Goodman's model for intensifiers

Based on the RSA model, Bennett & Goodman (2015) proposed a purely pragmatic account for degree modifiers. Though agreeing with Klein on that modified adjective phrases have the same semantics as unmodified ones except for a threshold shift, they claimed that intensifiers such as *very* or *extremely* give rise to the threshold shift in a non-compositional way, by simply changing the cost function Cost(u). To be concrete, modified adjective phrases such as *very tall* and *extremely tall* have the same semantics as the plain *tall*, except for different context-dependent thresholds. Since *extremely* is more costly than *very*, the threshold for *extremely tall* is higher than that of *very tall*. The cost parameters $c_{\text{rsa}}/c_{\text{som}}$ may depend on the length of the intensifier (longer words cost more than shorter ones) and the frequency (rarer words are harder to access, hence also cost more), etc. This account can be easily implemented using the RSA model, and Bennett & Goodman have already made predictions that match experimental data. In this paper, we want to further compare their account with Klein, by implementing both of them in both SOM and RSA. We hope to get insights into the two Bayesian models during the process.

1.4 Arguments for and against each Account

Before proceeding to the project, it seems necessary to discuss some theoretical arguments for or against each of the two accounts of degree modifiers. The key debate, as indicated above, lies on whether the intensifiers contribute to the threshold-shift compositionally, with non-vacuous lexical semantics. As Bennett & Goodman pointed out, though it is intuitive to encode the strengths of intensification into the lexical meanings of degree modifiers, it faces certain obstacles. First and foremost, there is a large multitude of degree modifiers and great potential for language production. For example, adverbs like *ridiculously* normally do not indicate an intensifying reading, but when used in *ridiculously tall*, we can easily construe it as an intensification of *tall*. In this sense, to provide lexical semantics for each intensifier would greatly affect theoretical parsimony.

On the other hand, Bennett & Goodman's account also suffers certain deficiencies. For one thing, it cannot exclude the possibility that the cost induced by an intensifier has something to do with its lexical meaning. As is mentioned in §1.2, the cost of an intensifier may depend on its length and frequency, but it is reasonable to argue that the word is rarely used because of its relatively extreme

meaning. Hence before we accept this account, we need to gain more evidence regarding the causal direction. Moreover, since the account is purely pragmatic and non-compositional, it faces direct objections with respect to compositionality. Consider the following sentences.

(10) a. Ronald is not *extremely tall*.
b. Ronald is *extremely tall* and *quite smart*.

Intuitively, (10a) means Ronald's height doesn't exceed the average height saliently, but may indicate that he can be relatively tall (or serves as a euphemism to say he's not tall). However, with Bennett & Goodman's account, the intensifier *extremely* doesn't contribute to the semantic meaning at all. But since it significantly strengthened the meaning of the adjectival phrase *not tall* (with a high cost), we can derive the meaning that 'Ronald is extremely not tall', which contradicts the general intuition that *extremely* is interpreted in the scope of negation. Also, (10b) means that Ronald's height saliently exceed the average, and his intelligence is somewhat above average. However, if we construe the intensification as purely pragmatic, we lose the binding between *extremely* and *tall* as well as *quite* and *smart*, and fail to derive the correct reading.

This paper doesn't have preference for either account. Rather, we hope to provide empirical evidence for/against them, by implementing them respectively with SOM and RSA, and compare their predictions with experimental data. The project will be introduced in detail in §2, with further discussions and future directions in section §2.4.

2 Project: Data and Results

2.1 Goal

The goal of this project is to test two different accounts proposed by K and Bennett & Goodman, respectively, for the interpretation of degree modifiers, embedded in both SOM and RSA. We expect it to provide empirical evidence for/against either of the interpretations, and to provide insights for the comparison between the two Bayesian models. Specifically, we want to see how well the quantitative predictions (from different models following different interpretations) fit with empirical data (obtained from the experiment conducted by Leffel et al., 2018).

2.2 Model and Data

Leffel et al. (2018) measured participants' agreement with sentences such as "Ronald is tall" given Ronald's exact height. They tested both 'tall' and 'very tall' (among other constructions), for 13 different heights from 5ft3in (160cm) to 6ft10in (208cm). Participants adjusted a slider to express their agreement. For our purpose, we interpret these judgments as reflecting the probability that

a sentence is true, i.e. the probability that $\theta \leq h$.[1] For SOM, this translates naturally as the cumulative distribution function of θ, $\int_{-\infty}^{h} Pr(\theta)d\theta$. For RSA, we will translate this as the posterior cumulative distribution of θ, as inferred by the pragmatic listener: $\int_{-\infty}^{h} \rho(\theta|u_1)d\theta$.

We tested all combinations of the two theoretical claims (Klein vs. Bennett & Goodman) and two probabilistic models (SOM vs. RSA), against the median judgment for each point of the scale and each construction. While in principle the models make predictions about individual speakers rather than a population (particularly the RSA), we chose to model the latter for simplicity and because the individual data was rather noisy. The median was preferred to the mean, as it is less sensitive to outliers and because the mean would not converge to 0% or 100% for extreme values. In each case, we started by adjusting the model parameters (cost c for *tall*, λ, prior on heights) to fit the data for *tall*, and then evaluated the best possible fit for *very tall*. All results are presented in Fig. 1.

(i) **Klein + SOM**
To fit the data on *tall*, we chose a normal prior distribution for $\phi(h)$ with parameters $\mu = 68.5$in, $\sigma = 3.7$in, the degree of rationality was $\lambda = 1.2$, and the cost for *tall* was $c_{\text{som}} = 0.2$.
According to Klein's account for degree modifiers, intensified adjectives such as *very tall* are interpreted as *tall compared to the set of tall people*. It can be incorporated into SOM by using the posterior distribution on heights after an utterance of *tall* as the prior distribution $\phi'(h)$ for the height of *tall people*:

$$\phi'(h) = \phi(h|\theta, u_1) \propto \phi(h) \int_{-\infty}^{h} \frac{Pr(\theta)}{1-\Phi(\theta)} d\theta \quad (11)$$

To simplify the further computation, we approximated this distribution with a Gaussian in the next steps. Combining (7),(8),(9) and (11), we can derive the distribution of the threshold θ' for *very tall*, which can then be integrated to derive a model of participants' judgments as $P(\theta' < h)$. The cost parameter for this second iteration of the algorithm was c'_{som}, and was meant to reflect the cost of adding *very* to the sentence. With constraint $c'_{\text{som}} \geq 0$, the optimal choice ended up being 0.

(ii) **Bennett & Goodman + SOM**
According to Bennett & Goodman, intensifying degree adverbs shift the threshold just because they increase the cost of utterances. Therefore, we could translate this account into the language of SOM simply by increasing the value of the cost parameter to $c_{\text{som}} + c'_{\text{som}}$, where c'_{som} is the additional cost caused by *very*. Here the optimal choice was $c'_{\text{som}} = 1.8$. Other parameters were identical to what we used to implement K's account.

[1] Leffel et al. (2018) interpret their results as reflecting not just truth but also pragmatic felicity (i.e. truth of the sentence together with its implicatures). However the implicatures they discuss only surfaced for more complex sentences (involving negation), and should therefore not affect our interpretation of the simpler sentences discussed here.

(iii) **Klein + RSA**
To fit the data on *tall* with the RSA, we chose a normal prior distribution for $\phi(h)$ with parameters $\mu = 69$in, $\sigma = 3.7$in, the degree of rationality was $\lambda = 4.8$, and the cost for *tall* was $c_{\text{rsa}} = 0.85$.
The RSA naturally provides the posterior distribution $\rho(h|u_1)$ for the heights of *tall people*:

$$\phi'(h) = \rho(h|u_1) \propto \phi(h) \int_{-\infty}^{h} \frac{Pr(\theta)}{1+e^{\lambda c_{\text{rsa}} \cdot (\int_{\theta}^{\infty} \phi(h')dh')^{\lambda}}} d\theta \quad (12)$$

Then combining (12) with (4),(5),(6), we derive Klein's predictions for *very tall* within the RSA model. A cost $c'_{\text{rsa}} = 1/3 c_{\text{rsa}} = 0.28$ gave close to optimal results.

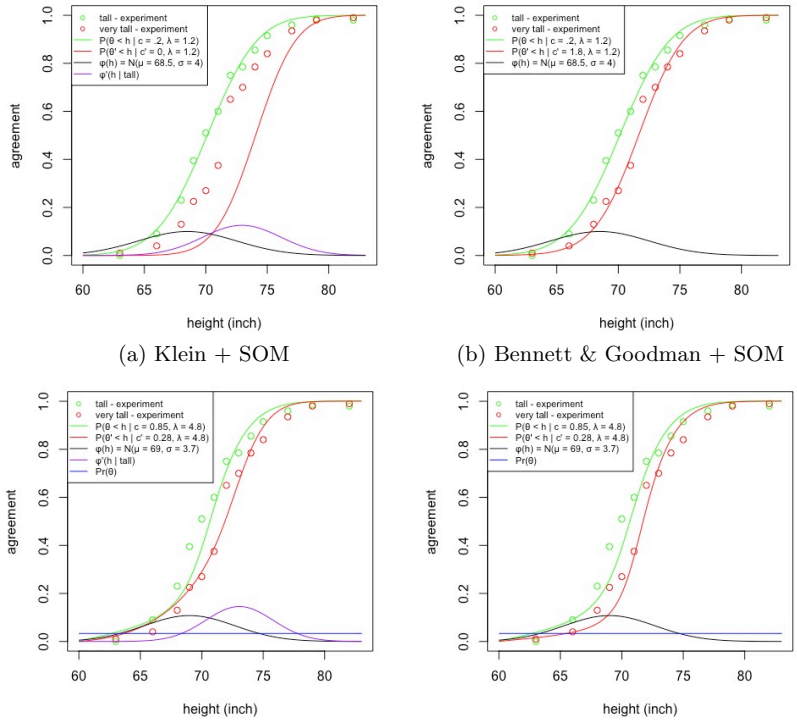
(iv) **Bennett & Goodman + RSA**
As in (ii), Bennett & Goodman's account of *very* only requires increasing the cost to $c_{\text{rsa}} + c'_{\text{rsa}}$. Setting all parameters as in (iii) gave good results.

2.3 Results

Fig. 1 shows the results of comparing empirical data with model predictions described above. Fig. 1a indicates that even with a minimal cost $c'_{\text{som}} = 0$ for the intensifier, the prediction of Klein's account implemented within SOM shifts the threshold of *very tall* far too right. All other three combinations gave good approximations of the data, although Bennett & Goodman's account within SOM required a very high cost $c'_{\text{som}} = 1.8$ (in comparison to $c_{\text{som}} = 0.2$ for the full unmodified sentence). Both accounts of *very* could be implemented in RSA with a reasonable cost for *very*.

2.4 Discussion

In Fig. (1a), we chose to present the radical case $c'_{\text{som}} = 0$ because any larger value of c'_{som} shifts the curve further right; and since this radical case has already yielded a result far to the right, it indicates that SOM + Klein cannot make correct predictions for *very tall*. Second, Fig. (1b), though exhibiting an accurate prediction, requires an additional cost parameter $c'_{\text{som}} = 1.8$. If we follow the assumption that cost is proportional to the length of expression as measured by the number of words (see Lassiter & Goodman, 2014), then compared to the cost $c_{\text{som}} = 0.2$ set to fit the data of *tall*, which is induced by the sentence *Ronald is tall*, c'_{som} is impractically large. In short, SOM's low sensitivity to the cost parameter (presented as an advantage of this model in Qing & Franke, 2014) may hinder itself in correctly predicting the meaning of degree modifiers. On the other hand, Fig. (1c) and (1d) show that RSA does give good predictions with both accounts, and with relatively practical cost parameters $c_{\text{rsa}} = 0.85$ and $c'_{\text{rsa}} \approx \frac{1}{3} c_{\text{rsa}}$. Note that our results offer an independent assessment of Bennett & Goodman (2015), as they only tested their models against a point-wise estimate of the posterior (but for multiple intensifiers), while we tested the model's



(a) Klein + SOM

(b) Bennett & Goodman + SOM

(c) Klein + RSA

(d) Bennett & Goodman + RSA

Fig. 1: Best fit for each model, compared to the median response in Leffel et al. (2018)

ability to fit the full posterior distribution (though only for *very*). Interestingly however, the present results do not clearly advocate between Klein and Bennett & Goodman.

3 Conclusion

We presented a four-way comparison between i) two theories of degree modifiers, i.e. Klein (1980) and Bennett & Goodman (2015), and ii) two recent probabilistic models for scalar adjectives, i.e. RSA (Lassiter & Goodman, 2014) and SOM (Qing & Franke, 2014), in dealing with degree modifiers. The comparison was conducted by testing the 2×2 combinations of theories and models on the experimental data from Leffel et al. (2018). The results showed us that SOM, due to its low sensitivity to the cost parameter, does not make a good prediction for the

meaning of intensifiers, whereas RSA presented us with relatively good results. Meanwhile, the results did not show preference between Klein and Bennet & Goodman. In particular, we could imagine implementing Bennett & Goodman's flexible account of various intensifiers by varying the cost in the second derivation involved in Klein's account (so *extremely tall* would also mean "tall among tall people", but with a higher cost than *very tall*). In fact, this is virtually the strategy we used trying to find the best fit for Klein's account with both RSA and SOM models.

In this paper, we only discussed the relative standard adjective *tall*. Turning to minimum-standard adjectives, such as *late* would in principle help further distinguish between the proposals of Klein and Bennett & Goodman. Crucially, the correct account should explain for the fact that *late* is minimum-standard while *very late* is relative. Nevertheless, neither the SOM nor the RSA was in position to fit the data for *late* presented in Leffel et al. (2018),[2] forcing us to leave this question for future research.

Acknowledgements

The author would like to thank Dr. Alexandre Cremers for supervising the project. I'm also grateful to the anonymous reviewers for their valuable feedback. Last but not least, many thanks to ESSLLI 2018 Grant Committee and EACL for the student grants.

Reference

Bennett, Erin & Noah D Goodman. 2015. Extremely costly intensifiers are stronger than quite costly ones. In Noelle et al (ed.), *Proceedings of the 37th annual meeting of the cognitive science society*, Austin, TX.

Burnett, Heather. 2014. A delineation solution to the puzzles of absolute adjectives. *Linguistics and philosophy* 37(1). 1–39.

Franke, Michael & Gerhard Jäger. 2016. Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft* 35(1). 3–44.

[2] One particularly difficult aspect of the data is that participants do not treat the predicate as pure minimum-standard, but still place a significant probability mass on the bottom of the scale. The distribution of θ is a mixed distribution with roughly 2/3 of the probability mass on the bottom point and the remaining 1/3 distributed continuously to the right of this point. This is in line with the more classical accounts of Burnett (2014) who argue that absolute adjectives show some tolerance, but only in one direction. The RSA model of Lassiter & Goodman (2015) cannot derive the discrete probability mass at the bottom point, while the SOM of Qing & Franke (2014) either derives a pure minimum-standard or a pure relative distribution (depending on how much probability mass of the prior is located at the bottom of the scale).

Kennedy, Christopher. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy* 30(1). 1–45.

Kennedy, Christopher & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 345–381.

Klein, Ewan. 1980. A semantics for positive and comparative adjectives. *Linguistics and philosophy* 4(1). 1–45.

Lassiter, Daniel & Noah D Goodman. 2014. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory*, 587–610.

Lassiter, Daniel & Noah D Goodman. 2015. Adjectival vagueness in a bayesian model of interpretation. *Synthese* .

Leffel, Timothy, Alexandre Cremers, Jacopo Romoli & Nicole Gotzner. 2018. Vagueness in implicature: The case of modified adjectives. Under Revision for *Journal of Semantics*.

Qing, Ciyang & Michael Franke. 2014. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Semantics and linguistic theory*, vol. 24, 23–41.

Disjunction under Deontic Modals: Experimental Data

Ying Liu

Utrecht University

Abstract. The meaning components of *may/or* and *must/or* sentences have been discussed intensively by a number of theoretical accounts. The debates are concerned with whether free choice inferences are logical interpretations or scalar implicatures, and whether exhaustive inferences and exclusive *or* inferences are derived for *may/or* versus *must/or*. In this study, two experiments were conducted to evaluate the assumptions of three representative accounts, namely, Fox (2007), Geurts (2005) and Simons (2005). Each experiment separately examined the availability and processing time-course of the three types of inferences associated with *may/or* versus *must/or* sentences. The experimental results are consistent with Simons (2005) to a large extent.

1 Introduction

This experimental study has two focuses. First, it presents a set of contrastive data which illustrate how people interpret and process sentences with disjunction embedded under a deontic possibility modal as shown in (1a) versus a deontic necessity modal as shown in (1b). Second, it discusses to what extent different theoretical accounts explain the data by looking into their assumptions.

(1) a. Mary may eat an apple or a banana. ***may/or*** sentence
b. Mary must eat an apple or a banana. ***must/or*** sentence

The interpretation of *may/or* and *must/or* sentences is closely associated with three types of inferences: free choice inferences as shown in (2a), exhaustive inferences as shown in (2b), and exclusive *or* inferences as shown in (2c). Intuitively, it is relatively certain that both types of sentences yield strong free choice inferences; however, it is unclear whether there exists a difference between them in triggering exhaustive and exclusive *or* inferences.

(2) a. **free choice inference** (options indicated by each disjunct are permitted): Mary is permitted to eat an apple, and she is also permitted to eat a banana.
b. **exhaustive inference** (options not indicated by disjuncts are not permitted): Mary is not permitted to eat anything other than an apple or a banana.
c. **exclusive *or* inference** (only one option is permitted at a time): Mary is not permitted to eat both an apple and a banana.

Relevant theoretical discussions (see [1], [3], [4], [9], [10], [11], [15], [17], [19]) primarily center on free choice inferences because free choice inferences are special in the sense that the standard semantics (i.e., the combination of the Boolean analysis of disjunction [2] and the standard modal logic for deontic modals [12]) completely fails to explain them, while the standard Neo-Gricean reasoning [15] can only account for their derivation for *must/or* sentences. In order to universally explain free choice inferences, in general, two types of accounts are developed: scalar implicature accounts (e.g. [3], [9]), and semantic accounts (e.g. [10], [15], [19]). It is impossible to tell which (types of) account is on the right track based on previous experimental studies (e.g. [6], [7], [18]) because they limit their investigations to free choice inferences drawn from *may/or* sentences, while exhaustive and exclusive *or* inferences, as well as the interpretation of *must/or* sentences, are not investigated.

In this study, I create a lottery machine paradigm, which separately examines the availability and derivation mechanism of the three types of inferences drawn from *may/or* versus *must/or* sentences. Based on the novel data, I attempt to evaluate different accounts. The study is structured as follows. In Section 2, I discuss the assumptions of three representative accounts, namely, Fox (2007) [9], Geurts (2005) [10] and Simons (2005) [15], and how they are experimentally testable. Section 3 and 4 separately report the experimental investigations on *may/or* and *must/or* sentences. In Section 5, I compare the data of *may/or* sentences with those of *must/or* sentences and discuss how the data may shed lights on theoretical accounts.

2 Background

Scalar Implicature Account Fox (2007) proposes that there exists a covert exhaustification operator (Exh) which can optionally be applied to *may/or* and *must/or* sentences, so that free choice inferences are explained as being derived by the same mechanism as that for scalar implicatures.

Exh needs to operate over a formally defined set of alternatives that is semantically closed under disjunction. For example, the set of alternatives of (1a) is closed as *Mary may eat an apple or a banana*, *Mary may eat an apple*, *Mary may eat a banana*, *Mary may eat an apple and a banana*. Only propositions that belong to the closed set of alternatives are relevant to the question under discussion and can be updated to the context set, while all other propositions are irrelevant and should be excluded (see Stalnaker (1978) [16]). As a result, exhaustive inferences are derived because the type of worlds in which Mary eats a thing other than an apple and a banana are excluded from the set of worlds of evaluation of (1a). Due to the same reason, exhaustive inferences are also assumed to be present for *must/or* sentences.

One primary role of Exh is to eliminate as many alternatives as possible. When Exh is applied to a *may/or* sentence such as (1a) for the first time, the stronger alternative containing and (e.g. *Mary may eat an apple and a banana*) is negated, and an exclusive *or* inference is derived. After the first-step exhaus-

tification, a new set of alternatives is generated, which includes *Mary may eat an apple but not a banana* and *Mary may eat a banana but not an apple*. These alternatives can simultaneously be negated if Exh is further applied to the sentence. The recursive application of Exh eventually gives rise to a free choice inference.

Different from *may/or* sentences, a *must/or* sentence such as (1b) only involves a one-step exhaustification. Once Exh is applied, the simultaneous negation of the stronger alternatives containing individual disjuncts (e.g. *Mary must eat an apple* and *Mary must eat a banana*) gives rise to a free choice inference, while the negation of the stronger alternative containing and (e.g. *Mary must eat an apple and a banana*) gives rise to an inference implying that it is not obligatory for Mary to eat an apple and a banana. This inference is compatible with the type of worlds in which Mary is permitted to eat an apple and a banana, and she can also freely choose whether to eat them. Thus, exclusive *or* inferences are not expected for *must/or* sentences.

Semantic Accounts Geurts (2005) and Simons (2005) solve the free choice puzzle by proposing alternative semantics. Their accounts are crucially different from each other in how they deal with the scope relation between disjunction and deontic modals and how they formulate the semantics for disjunction. More specifically, Geurts claims that if assuming that disjunction takes scope over deontic modals, sentences with disjunction embedded under deontic modals can be analyzed as conjunctions of modal propositions. In comparison, Simons proposes that under the scope of deontic modals, disjunction can introduce sets of alternative propositions.

Despite their differences, they similarly argue that free choice inferences are derived as the results of the computation of truth conditions, and they are the preferred logical interpretations of both *may/or* and *must/or* sentences. In addition, they also similarly suggest that there is some kind of exclusive *or* constraint [1] which can be applied to *may/or* and *must/or* sentences to restrict the intersection between the sets of worlds denoted by individual disjuncts. They further suggest that this constraint is a pragmatic constraint, and the exclusive *or* inferences triggered by it is sort of conversational implicatures.

Although Geurts and Simons hold the same point of view that the exhaustive effect should always be a semantic effect, they have noticeably different assumptions concerning whether the effect should be available for *may/or* versus *must/or* sentences. According to Geurts who proposes that disjunction takes a wider scope, the existence of the exhaustive effect is completely dependent on whether disjunction should be closed. Following Zimmermann (2000) [19], he claims that the conjunctive lists of modal propositions coordinated by disjunction should be closed by default unless they are explicitly marked by intonation or other linguistic devices. Therefore, usually, exhaustive inferences are derived semantically for both *may/or* and *must/or* sentences. Based on Simons who suggests that deontic modals take a wider scope, the semantics of deontic modals

[1] The exclusive *or* constraint is named as the disjointness constraint in Geurts (2005, p. 395) and the no total overlap constraint in Simons (2005, p. 29).

plays a central role in determining whether the exhaustive effect is present. Exhaustive inferences are present for *must/or* sentences because must, as a deontic necessity modal, prohibits the deontic accessibility to any type of worlds that are not denoted by individual disjuncts. In comparison, exhaustive inferences can be absent for *may/or* sentences because may, as a deontic possibility modal, allows for the accessibility to an arbitrary type of worlds as long as the types of worlds denoted by individual disjuncts are deontically accessible.

Psychological Implications Studies on language processing (e.g. [5], [8], [13])[2] suggest that the computation of logical interpretations take place alongside with the computation of logical forms in syntax, so it is automatic with low cognitive costs involved. In comparison, scalar implicature derivation is cognitively costly, because addressees may need spend cognitive resources in reasoning why addressers utter a specific scalar expression instead of its stronger alternatives or deciding whether a deep reasoning process should be applied. Since cognitive resources in working memory are limited, it is fairly difficult for people to derive scalar implicatures for every single occurrence of scalar expressions. As a result, scalar implicatures are only optionally derived.

In psycholinguistic experiments, the degree of availability of an inference, reflected by derivation rates, conveys information about whether this inference is computed automatically or optionally; while the processing time conveys information about whether an inference is cognitively costly. Based on these, if an inference is derived semantically, the derivation rate of it should be very high (i.e., ideally close to 100%), and the processing time of it should be similar as that of a logical interpretation. By contrast, if an inference is a scalar implicature in nature, the derivation rate of it should be moderate (i.e., a value that is neither close to 0% nor 100%), and the processing time of it should be much longer than that of a logical interpretation. Thus, by experimentally examining the derivation rate and processing time, we can tell whether an inference is available and whether it is a logical interpretation or a scalar implicature.

In short, once we obtain the derivation rate and processing time of each of the three types of inferences associated with *may/or* and *must/or* sentences, we can immediately know the meaning components of *may/or* versus *must/or* sentences and the nature of these components. Based on these, we can further judge the plausibility of different theoretical accounts.

3 Experiment for *May/or* Sentences

Purposes To examine derivation rates and processing time of free choice, exhaustive and exclusive *or* inferences triggered by *may/or* sentences.

Participants 40 Dutch native speakers (aged 18 and above) were recruited from the Dutch participant database of Utrecht University.

[2] I only discuss the studies which adopt a method similar as the one I used, i.e., a picture-sentence binary judgment task, because in psycholinguistic experiments, differences in the types of tasks involved can cause crucial differences in results.

Method I designed a lottery game paradigm in Dutch in ZEP [3], which required participants to do an online picture-sentence binary judgment task based on the cover story as shown in Figure 1. The cover story was only presented once at the beginning of the task. Participants were asked to read the cover story carefully without a time constraint, and their understandings about the story were examined by 6 practice trials. Only when they successfully passed all practice trials, they could start the test session.

The picture shows a lottery machine for children. After a child wins a lottery game, the small screen on the right hand side of the machine will display the total amount of cash prize the child has been awarded. The central screen of the machine will display six different items. The price and availability of the items are displayed below each one of the items. The price is indicated in Euros. A green light indicates an available item and a red light indicates that the item is not available for purchase. The price of the items and their availability as well as the amount of cash prize the child has been awarded all determine what items the child is allowed to buy from the lottery machine.

Fig. 1. English Translation of the Cover Story for *May/or* Experiment

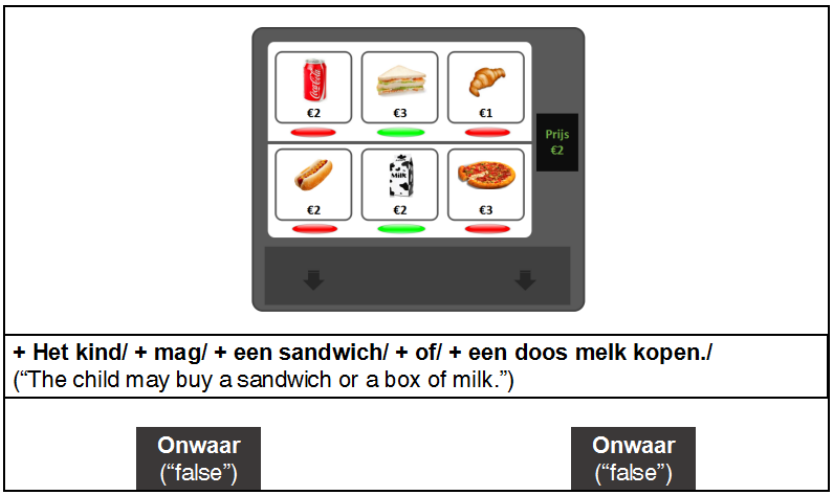


Fig. 2. Example of Target Trial in *May/or* Experiment

Figure 2 illustrates one target trial. For each trial, a picture depicting a lottery machine would firstly be presented on computer screen for 500ms. After this,

[3] Information about ZEP can be found at https://www.beexy.nl/

a plus sign (+) would occur at the beginning of the sentence bar beneath the picture. Participants were instructed that if they press the middle button on the button box in front of them, the plus sign would be replaced by the first chunk of the sentence. The sentence chunks could continuously show up by pressing the same button. Participants could read the sentence at their own pace by controlling the speed of button press. All *may/or* sentences were cut into five chunks as shown in Figure 2. Once the entire sentence was presented on the screen, participants were required to judge whether the sentence is true or false based on the cover story by pressing the corresponding left/right button on the button box. *True/false* responses and the reaction time from the occurrence of the last sentence chunk to the left/right button press were recorded.

The *False* response in Figure 2 indicates the existence of the free choice inference. To illustrate, based on the cover story, the picture indicates that the only item the child is permitted to buy is a box of milk because it is both available and affordable. If participants derive the free choice inference for the sentence, which implies that the child is permitted to buy a sandwich and he/she is also permitted to buy a box of milk, they should judge the sentence as the incorrect description of the picture. Thus, the percentage of *false* responses in trials like the one in Figure 2 indicates the derivation rates of free choice inferences, while the reaction time of these *false* responses reflects the processing time needed for deriving free choice inferences. By adopting a very similar design of trials, the data of exhaustive and exclusive *or* inferences can also be obtained.

Design and Materials A single factor within-subject design was used. The independent variable was the type of conditions created for *may/or* sentences. Each *may/or* sentence occurred in three target conditions (i.e., the free choice, exhaustive and exclusive *or* conditions) and two control conditions in which only logical interpretations were involved in making judgments (see Figure 3). There were two dependent variables: the type of responses (i.e., *true* or *false*) and the reaction time associated with the responses. 12 trials were created for each target/control condition. 70 filler trials were added to conceal experimental purposes. All trials were pseudo-randomized.

Predictions Fox (2007) predicts that exhaustive inferences are derived semantically, while free choice and exclusive *or* inferences are derived as scalar implicatures. Based on this, the percentage of *false* responses in the exhaustive condition should be near 100%, and the corresponding reaction time should be roughly the same as that in control conditions. The percentage of *false* responses in the free choice and exclusive *or* condition should be neither close to 0% nor 100%, and the corresponding reaction time should be longer than that in control conditions.

Geurts (2005) predicts that exhaustive and free choice inferences are derived semantically, while exclusive *or* inferences are derived as conversational implicatures. Based on this, the percentage of *false* responses in the free choice and exhaustive condition should be near 100%, and the corresponding reaction time should be roughly the same as that in control conditions. The percentage of *false* responses in the exclusive *or* condition should be neither close to 0% nor

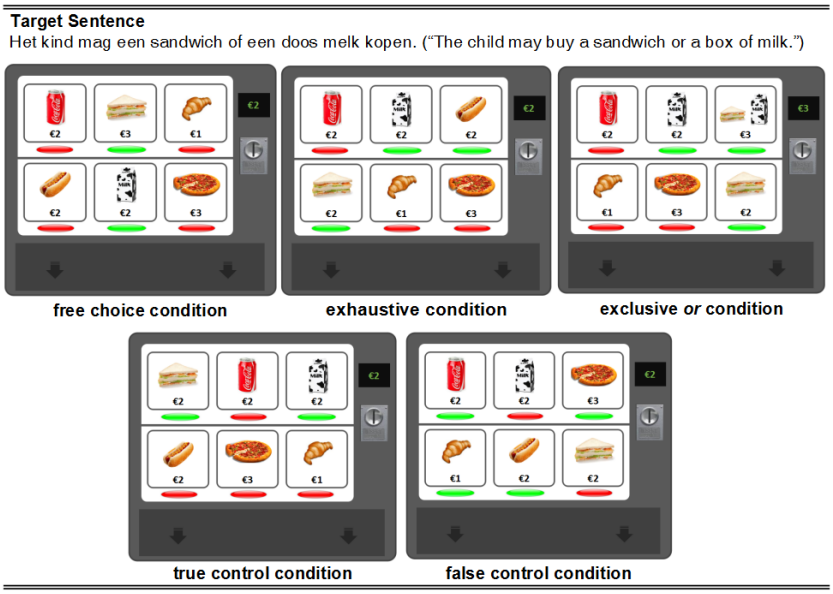


Fig. 3. Target and Control Conditions in *May/or* Experiment

100%, and the corresponding reaction time should be longer than that in control conditions.

Simons (2004) predicts that free choice inferences are derived semantically and exclusive *or* inferences are derived as conversational implicatures, while exhaustive inferences are absent. Based on this, the percentage of *false* responses in the free choice condition should be near 100%, and the corresponding reaction time should be roughly the same as that in control conditions. The percentage of *false* responses in the exclusive *or* condition should be neither close to 0% nor 100%, and the corresponding reaction time should be longer than that in control conditions. The percentage of *false* responses in the exhaustive condition should be near 0%.

Results I excluded the data from 7 participants whose accuracy rates on control items were low. I removed all reaction time data associated with wrong responses in control conditions, and I further removed 5.78% reaction time data which are outliers.[4]

The percentages of *false* responses, indicating the derivation rates of the inferences under investigation, are given in Figure 4. The derivation rate of free choice inferences is as high as 92.42%. The derivation rates of exhaustive and exclusive *or* inferences were much lower, i.e., 15.15% and 28.54% respectively. The percentage of *false* responses in the free choice condition was significantly

[4] Outliers are more than 1.5 IQRs below the first quartile or above the third quartile.

higher than that in the true control condition (= 10.77, SE = 0.66, z = 16.19, p < 0.001), and it was significantly lower than that in the false control condition (= -2.36, SE = 0.57, z = -4.13, p < 0.001). The percentage of *false* responses in the exhaustivity condition was significantly higher than that in the true control condition (= 3.49, SE = 0.53, z = 6.60, p < 0.001), and it was significantly lower than that in the false control condition (= -9.63, SE = 0.68, z = -14.16 , p < 0.001). The percentage of *false* responses in the exclusive *or* condition was significantly higher than that in the true control condition (= 5.23, SE = 0.56 , z = 9.35, p < 0.001), and it was significantly lower than that in the false control condition (= -7.90 , SE = 0.63, z = -12.50, p < 0.001). [5] The reaction time, reflecting the processing time of the inferences under investigation, is illustrated in Figure 5. The reaction time of free choice inferences (M ≈ 4724ms, SD ≈ 2228ms) were not significantly different from that in the true and false control condition. The reaction time associated with exhaustive inferences (M ≈ 6551ms, SD ≈ 2367ms) was significantly longer than that in the false control condition (= 9.60, SE = 2.04, t = 4.68, p < 0.001) and the true control condition (= 11.8, SE = 2.04, t = 5.78, p < 0.001). The reaction time associated with exclusive *or* inferences (M ≈ 5968ms, SD ≈ 2680ms) was significantly longer than that in the false control condition (= 7.10, SE = 1.61, t = 4.42, p < 0.001) and the true control condition (= 9.30, SE = 1.60, t = 5.81, p < 0.001). [6]

Discussion There are two important findings. First, free choice inferences

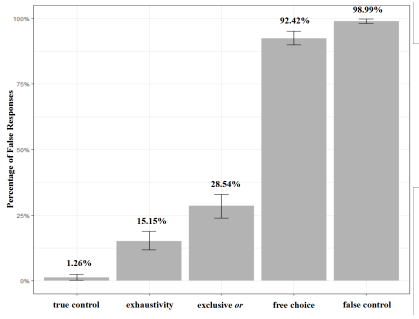


Fig. 4. *False* Responses (*May/or*)

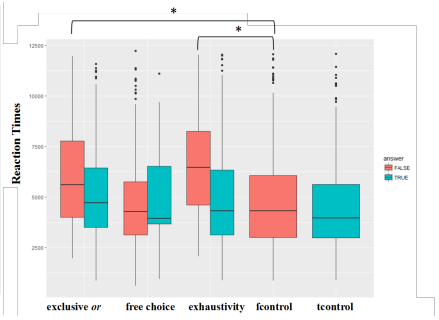


Fig. 5. Reaction Time (*May/or*)

were derived almost by default, and the derivation of them was not more time-consuming than that of logical meanings. Second, both exhaustive and exclusive

[5] The response data were submitted to a generalized linear mixed effects model in R (using the *glmer* function) with sentences, pictures and participants as randomized effects, and conditions as fixed effects.

[6] The reaction time data were submitted to a linear mixed-effects model in R (using the *lmer* function) with sentences, pictures and participants as randomized effects, and responses and conditions as fixed effects.

or inferences were only derived optionally, with the processing time significantly longer than that of logical meanings. Given these, free choice inferences are most likely to be the preferred logical interpretations of *may/or* sentences, while exhaustive and exclusive *or* inferences are most likely to be some sort of conversational implicatures if they indeed are not scalar implicatures.

Fox (2007) account has the poorest fit to the data because it completely fails to predict the derivation pattern of free choice and exhaustive inferences. Geurts (2005) faces a fatal problem in explaining the extremely low derivation rate of exhaustive inferences. Loosely speaking, Simonss (2004) account has a considerably good fit to the data. However, it is a bit surprising to find that exhaustive inferences were very occasionally derived as a kind of conversational implicatures for *may/or* sentences. As far as I am concerned, no theoretical study has ever considered this as a possibility.

4 Experiment for *Must/or* Sentences

Purposes To examine derivation rates and processing time of free choice, exhaustive and exclusive *or* inferences triggered by *must/or* sentences.

Participants 25 Dutch native speakers (aged 18 and above) were recruited from the Dutch participant database of Utrecht University.

Methods, design and materials The paradigm, set-up, design and materials of this experiment were exactly the same as those of the *may/or* experiment (see Section 3) except for two aspects. First, one piece of information was added to the cover story of this experiment, which stated that the child has to buy something from the lottery machine with the cash prize he/she has been awarded, otherwise the machine will be unable to load the next lottery game. Second, *moeten* (must) instead of *mogen* (may) was used in all target sentences.

Predictions Fox (2007) predicts that exhaustive inferences are derived semantically and free choice inferences are derived as scalar implicatures, while exclusive *or* inferences are not derived. Based on this, the percentage of *false* responses in the exhaustive condition should be near 100%, and the corresponding reaction time should be roughly the same as that in control conditions. The percentage of *false* responses in the free choice condition should be should be neither close to 0% nor 100%, and the corresponding reaction time should be longer than that in control conditions. The percentage of *false* responses in the exclusive *or* condition should be near 0%.

Geurts (2005) and Simons (2004) predict that exhaustive and free choice inferences are derived semantically, while exclusive *or* inferences are derived as conversational implicatures. Based on this, the percentage of false responses in the free choice and exhaustive condition should be near 100%, and the corresponding reaction time should be roughly the same as that in control conditions. The percentage of *false* responses in the exclusive *or* condition should be neither close to 0% nor 100%, and the corresponding reaction time should be longer than that in control conditions. Please especially notice that Geurtss (2005) predictions for *may/or* and *must/or* sentences are exactly the same.

Results I included all participants response data. I removed all reaction time data associated with wrong responses in control conditions, and I further removed 5.47% reaction time data which are outliers.

According to Figure 6, the derivation rates of free choice, exhaustive and

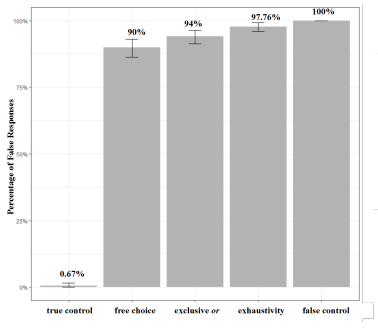


Fig. 6. *False* Responses (*Must/or*)

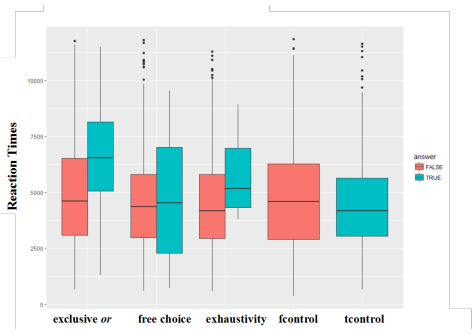


Fig. 7. Reaction Time (*Must/or*)

exclusive *or* inferences associated with *must/or* sentences were all very high, i.e., 90%, 97.76% and 94% respectively. The percentages of *false* responses in all three target conditions were only significantly different from that in the true control condition, but they were not significantly different from that in the false control condition. More specifically, the percentage of *false* responses in the false control condition was not significantly higher than that in the free choice condition (= 19.17, SE = 2284.02, z = 0.01, p ≈ 0.99), the exclusive or condition (= 18.45, SE = 2284.02, z = 0.01, p ≈ 0.99), and the exhaustivity condition ((= 17.15, SE = 2284.02, z = 0.01, p ≈ 0.99).

Figure 7 summarizes the reaction time data. The reaction time of free choice inferences (M ≈ 4715ms, SD ≈ 2387ms), exhaustive inferences (M ≈ 4659ms, SD ≈ 2326ms) and exclusive or inferences (M ≈ 5013ms, SD ≈ 2594ms) were all not significantly different from that of the true and false control condition.

Discussion The main finding is that all three types of inferences were found to be derived by default for *must/or* sentences, with the processing time not significantly different from that of logical interpretations in control conditions. It seems that all three types of inferences are parts of the preferred logical interpretations of *must/or* sentences. Foxs (2007) account still only very poorly fits the data. It only predicts the derivation pattern of exhaustive inferences. Both Geurtss (2005) and Simonss (2004) accounts successfully explain the data associated with free choice and exhaustive inferences; however, none of them predicts the default and rapid derivation of exclusive *or* inferences.

Concerning the data associated with exclusive *or* inferences, I think there are two possibilities. First, it is possible that due to the existence of the deontic necessity modal, the permissions expressed by *must/or* sentences are actually

much stronger than we thought. So the exclusive *or* constraint might not be a pragmatic constraint but a construction-specific semantic constraint. However, this possibility does not seem to be very plausible because the construction-specific assumption violates the principle of least effort. Second, it is possible that this piece of data is not sufficiently valid due to the potential problem in the experimental paradigm. To illustrate, in the cover story, we added one piece of information which stated that the child has to buy something from the lottery machine. It could be possible that a large number of participants understood something as denoting exactly one instead of at least one undermined thing. As a result, they might automatically ruled out the possibility that the child is permitted to buy two things at once. More experiments need be done to better understand the nature of exclusive or inferences triggered by *must/or* sentences.

5 General Discussion

If comparing the data of *may/or* with those of *must/or* sentences, one similarity and two differences can be found. The similarity is that free choice inferences are the preferred logical interpretations of both *may/or* and *must/or* sentences. So generally, the semantic accounts, i.e., Geurts (2005) and Simons (2004), are more plausible than the scalar implicature account, i.e., Fox (2007).

The crucial differences are that while exhaustive and exclusive *or* inferences were only very occasionally derived for *may/or* sentences as conversational implicatures, they were derived by default for *must/or* sentences as parts of logical interpretations. Lets temporarily not discuss the difference in exclusive *or* inferences (because no definite answer on the nature of exclusive *or* inferences of *must/or* sentences can be given), and only look into the difference in exhaustive inferences triggered by the two types of sentences. Simons (2004) successfully predicts this difference because she assumes that the alternative semantics for disjunction is activated under the scope of deontic modals. Due to this, even if a closure operation is applied to disjunction, the semantics of the deontic possibility modal still opens up the possibility for making *may/or* sentences non-exhaustive. Geurts (2005) fails to do so because he claims that disjunction takes scope over deontic modals, and in addition, it introduces a closed set of propositions. Once disjunction is closed, *may/or* sentences can only be exhaustive. Thus, Simonss (2004) account has the highest explanatory power.

To conclude, this study intends to convey three pieces of information: first, free choice puzzles are better to be solved semantically; second, differences in deontic modals may lead to differences in exhaustive and exclusive *or* inferences; third, the scope relation between disjunction and deontical modals should be dealt with carefully.

6 Acknowledgements

My thanks go to Yaron McNabb, Rick Nouwen and Henriette de Swart for discussions and suggestions and Chris van Run for the ZEP script.

References

1. Aloni, M.: Free choice, modals, and imperatives. *Natural Language Semantics*, *15*(1), 65-94 (2006)
2. Aloni, M.: "Disjunction". In: *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), ¡https://plato.stanford.edu/archives/win2016/entries/disjunction/¿ (2016)
3. Alonso-Ovalle, L.: *Disjunction in alternative semantics*. Doctoral dissertation, University of Massachusetts Amherst (2006)
4. Barker, C.: Free choice permission as resource-sensitive reasoning. *Semantics and Pragmatics*, *3*, 10-1 (2010)
5. Bott, L., Noveck, I. A.: Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, *51*(3), 437-457 (2004)
6. Chemla, E.: Universal implicatures and free choice effects: Experimental data. *Semantics and Pragmatics*, *2*, 2-1 (2009)
7. Chemla, E., Bott, L.: Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition*, *130*(3), 380-396 (2014)
8. Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., Sperber, D.: Making disjunctions exclusive. *The quarterly journal of experimental psychology*, *61*(11), 1741-1760 (2008)
9. Fox, D.: Free choice and the theory of scalar implicatures. In: *Presupposition and implicature in compositional semantics*, pp. 71-120. Palgrave Macmillan UK (2007)
10. Geurts, B.: Entertaining alternatives: Disjunctions as modals. *Natural language semantics*, *13*(4), 383-410 (2005)
11. Kaufmann, M.: Free choice is a form of dependence. *Natural Language Semantics*, *24*(3), 247-290 (2016)
12. Kratzer, A.: Modality. In: *Semantics: An international handbook of contemporary research*, ed. by Arnim von Stechow and Dieter Wunderlich, 63950 (1991)
13. Marty, P. P., Chemla, E.: Scalar implicatures: working memory and a comparison with *only*. *Frontiers in psychology*, *4* (2013)
14. Sauerland, U.: Scalar implicatures in complex sentences. *Linguistics and philosophy*, *27*(3), 367-391 (2004)
15. Simons, M.: Dividing Things Up: The Semantics of Or and the Modal. (2004)
16. Stalnaker, R. C. *Assertion*. Blackwell Publishers Ltd, pp. 147-161 (1978)
17. Starr, W.: Expressing permission. In *Semantics and Linguistic Theory*, *26*, pp. 325-349 (October 2016)
18. van Tiel, B.: Universal free choice. In *Proceedings of Sinn und Bedeutung*, *16*, pp. 627-638 (June 2012).
19. Zimmermann, T. E.: Free choice disjunction and epistemic possibility. *Natural language semantics*, *8*(4), 255-290 (2000)

"First things first": an Inquisitive Plausibility-Urgency Model

Zhuoye Zhao and Paul Seip

ILLC, University of Amsterdam

Abstract. There is a fruitful line of work in incorporating questions into epistemic logic (Van Benthem & Minică, 2009; Baltag et al., 2016, among others). Based on the viewpoint that communication is a process of raising and resolving issues, inquisitive semantics introduces a uniform notion of meaning for statements and questions, thus can serve as a suitable device for this purpose. For instance, Inquisitive Plausibility Model (Ciardelli & Roelofsen, 2014) is able to combine questions with the Epistemic Plausibility Model (IPM) (Baltag & Smets, 2006a,b) to capture not only the *belief* and *knowledge* of agents, but also the *issues* they entertain. Building on this, we develop an Inquisitive Plausibility-Urgency Model (IPUM), which not only allows us to model *knowledge*, *belief* and *issues*, but also the *urgency* of the *issues*, hence lead us to towards formalizations of more dynamics of questions.

1 Introduction

Classical models of epistemic change/belief revision (Van Ditmarsch et al., 2007; Van Benthem, 2007, among others) encode the knowledge/belief of an agent as a set of (non-inquisitive) propositions, whose semantics are often modeled as a set of possible worlds. One of the representatives is the Epistemic Plausibility Model(Baltag & Smets, 2006a,b):

Definition 1. An Epistemic Plausibility Model $\mathcal{M}$ for a set of agents $\mathcal{A}$ is a tuple:

$$\mathcal{M} = \langle W, \{\leq_a\}_{a\in\mathcal{A}}, \{\sigma_a\}_{a\in\mathcal{A}}, \|\cdot\|\rangle$$

where

- W is a set of possible worlds
- $\leq_a \subseteq W \times W$ encodes the plausibility map for each agent, as a converse well-founded total preorder between possible worlds
- σ_a is an epistemic map for each agent $a \in \mathcal{A}$: for every $w \in W$, $\sigma_a(w)$ is the epistemic state of a at w
- $\|\cdot\|$ is the valuation function

The knowledge/belief modalities (including knowledge K_a, belief B_a, strong belief Sb_a and conditional belief B_a^Q) can be semantically characterized as follows:

- $\mathcal{M}, w \vDash K_a P \Longleftrightarrow \sigma_a(w) \subseteq \| P \|$
- $\mathcal{M}, w \vDash B_a P \Longleftrightarrow Max_{\leq_a}\{w \in W : w \in \sigma_a(w)\} \subseteq \| P \|$
  (Henceforth, we abbreviate $Max_{\leq_a}\{w \in W : w \in \sigma_a(w)\}$ as $bel_a(w)$)
- $\mathcal{M}, w \vDash Sb_a P \Longleftrightarrow s >_a t$ for every $s, t \in \sigma_a(w)$ where $s \in \| P \|, t \notin \| P \|$
- $\mathcal{M}, w \vDash B_a^Q P \Longleftrightarrow bel_a(w) \cap \| Q \| \subseteq \| P \|$

Crucially, the plausibility relation $\leq_a$ is defined in a *conditional* manner, namely, fixing any two worlds $w_1, w_2 \in W$, if the agent a thinks w_2 is at least as *plausible* as w_1, then there is $w_1 \leq_a w_2$, and vice versa. Also note that the notion of *plausibility* indicates a certain degree of epistemic indistinguishability between different worlds, therefore it fully describes the epistemic state (all the worlds the agent thinks possible). Formally, for any $w' \in W$, $w' \in \sigma_a(w)$ iff $w' \leq_a w$ or $w \leq_a w'$. However, we will keep the notion of epistemic state here for convenience. Then for each agent a, her *Knowledge* at w is captured by the epistemic state $\sigma_a(w)$, meaning "all the possible worlds a knows for sure (that's possible) at w", whereas the *Belief* is captured by $best\sigma_a(w)$, which is determined by $\sigma_a(w)$ and the plausibility relation $\leq_a$, meaning "the most plausible set of worlds in $\sigma_a(w)$". Based on this model, the *dynamics* of the agent's epistemic/doxastic state, namely *knowledge update* or *belief revision* are construed as model transformers that map the current plausibility model to a new one.

However, in modeling the (dynamic) doxastic state of an agent, her *questions*, or rather, the *issues* she entertains are also important. As pointed out by Schaffer (2005), *All knowledge involves a question; To know is to know the answer*. Following the spirit of the 'Socratic epistemilogy' initiated by Hintikka in the 1970's and later proposed in Hintikka (2007), there is a fruitful line of work in incorprating questions into epistemic/doxastic logic (Olsson & Westlund, 2006; Enqvist, 2010, among others). Providing a uniform notion for propositions and questions, Inquisitive Semantics (Ciardelli, 2009; Groenendijk & Roelofsen, 2009; Ciardelli et al., 2013) is thus able to enrich the classical models in this aspect. In this paper, we will base our work on the inquisitive plausibility model (IPM) (Ciardelli & Roelofsen, 2014), and further extend it with an urgency order defined as a converse-well-founded total preorder between *information states*, which are characterized as sets of possible worlds. In the rest of this section, we present technical details of the background frameworks mentioned above. We will specify our model in §2, and then show some applications in §3. We will conclude in §4.

1.1 Inquisitive Semantics (InqB) and Inquisitive Epistemic Logic (IEL)

Inquisitive semantics (InqB) starts from the observation that the primary function of natural language is to exchange information, thus motivates a notion of meaning that captures not only informative, but also inquisitive content. To achieve this, inquisitive semantics generalizes the meaning of a sentence as the *issue* it raises. An *issue* is a set of propositions, thus formalized as a set of "sets of possible worlds", namely *information states*. An *issue* can be represented as

a set of information states that resolve it. Notice that if a certain proposition p resolves an issue, then any stronger proposition q, that is $q \subseteq p$, also resolves the issue. Therefore, we can formalize the notion of *issue* as a non-empty, downward closed set of propositions. In possible world semantics, a proposition p is *true* in a world w just in case $w \in p$; in parallel, an issue P (represented as a set of sets of possible worlds) is *supported/resolved* by a proposition p just in case $p \in P$. Here and henceforth, we will denote both the *truth* relation and the *support* relation as '$\models$'.

The maximal elements of an issue P are referred to as its *alternatives*, written as $alt(P)$. A sentence is *inquisitive* if it has more than one alternative, and is *non-inquisitive* if it has only one alternative. The *information content* of an issue P, denoted by $|P|$ or info(P), is defined as $\bigcup P$, the union of elements in P.

The language of InqB is very much like propositional logic, with atomic formula $p, q...$, negation $\neg$, boolean connectives $\vee, \wedge, \rightarrow$ of similar accounts, except for two addtional projection operators, ! and ?, which are referred to as *non-inquisitive* and *non-informative operators*, respectively. The non-inquisitive operator ! maps an issue P to the power set of its informative content, i.e. $!P = \mathcal{P}(|P|)$, while the non-informative operator ? maps P to the disjunction of itself and its negation, i.e. $?P = P \cup S \backslash |P|$.

Based on InqB, Inquisitive Epistemic Logic (IEL) characterizes the basic epistemic notions in the same fashion. In order to picture epistemic concepts, IEL introduces two basic epistemic notions. The *epistemic state* of an agent a at a world w, written as $\sigma_a(w)$, consists of the worlds that a considers possible. The *inquisitive state*, written as $\Sigma_a(w)$, can be read as the issue that the agent a *entertains*. Therefore, for any information state $s \in \Sigma_a(w)$, s is a resolution to the issue that concerns a. Moreover, the epistemic state of an agent is always equivalent to the informative content of its inquisitive state, that is, $\sigma_a(w) = \bigcup \Sigma_a(w)$. With these basic notions, we can then introduce two basic epistemic modalities that we will operate on.

Definition 2. The knowledge modality K_a

$$w \vDash K_a\phi \Longleftrightarrow \sigma_a(w) \in \phi$$

That is, an agent *knows* a sentence ϕ if and only if the agent's epistemic state resolves the issue raised by the sentence. Similarly, we can define a modality E_a that pictures the issues that the agent *entertains*.

Definition 3. The Entertain modality E_a

$$w \vDash E_a\phi \Longleftrightarrow \forall s \in \Sigma_a(w), s \in \phi$$

1.2 Inquisitive Plausibility Model (IPM)

Based on classical plausibility model, InqB and observations made in Olsson & Westlund (2006); Enqvist (2010), Ciardelli & Roelofsen (2014) proposed a semantic framework known as inquisitive plausibility model to capture not only

the belief and knowledge of an agent, but also the issues an agent entertains, which can be viewed as her "long-term epistemic goals". Further, it can be used to model the *research agenda* of an agent, which is captured as the issues the agent entertains conditioning on her belief. The formalizations are as follows.

Definition 4. An inquisitive plausibility model for a set of agents $\mathcal{A}$ *is a tuple* $\langle W, V, \{\sigma_a\}_{a\in\mathcal{A}}, \{\leq_a\}_{a\in\mathcal{A}}, \{\Sigma_a\}_{a\in\mathcal{A}}\rangle$ that consists of:

- a set W of possible worlds.
- a valuation function V.
- an epistemic map σ_a for each agent.
- a plausibility map $\leq_a$ for each agent.
- an inquisitive map Σ_a for each agent $a \in \mathcal{A}$*:* for every $s \in S$, $\Sigma_a(s)$ is an issue over $\sigma_a(s)$.

The language and semantics of a corresponding inquisitive belief logic is then naturally adapted from IEL and classical plausibility model. In addition to the knowledge and belief modalities defined above, we further introduce the modalities of *entertaining* E_a, *conditional entertaining* E_a^Q, etc. Here we denote information states by α and issues by $\mu :=?\{\alpha_1, \ldots, \alpha_n\}$. Also, we will always assume that μ is in the minimal form, i.e. $\alpha_1, ..., \alpha_n$ are non-redundant alternatives, and μ is the downward closure of them.

Resolution

$M, w \vDash ?\{\alpha_1, \ldots, \alpha_n\} \Longleftrightarrow$ for some α_i, $M, w \vDash \alpha_i$ for every $w \in s$.

Truth Conditions

- $\mathcal{M}, w \vDash E_a\mu \Longleftrightarrow \forall t \in \Sigma_a(w), t \in \mu$
- $\mathcal{M}, w \vDash E_a^Q\mu \Longleftrightarrow \forall t \subseteq \| Q \|, t \in \Sigma_a(w) \Rightarrow t \in \mu$
- $\mathcal{M}, w \vDash E_a^B\mu \Longleftrightarrow \forall t \subseteq bel_a(w), t \in \Sigma_a(w) \Rightarrow t \in \mu$

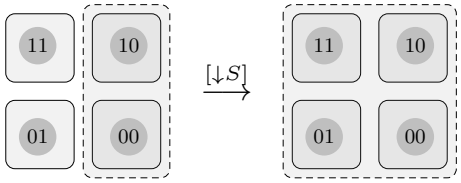
Note that E_a^B is the entertain-over-belief modality, which is used to address issues the agent is entertaining over her beliefs, i.e. the *research agenda*.

1.3 Inquisitive Contractions

Here we introduce an application of IPM. Classical belief revisions are captured as model transformations between plausibility models. Using IPM, we can not only preserve the classical operations, but also successfully model changes of an agent's research agenda. One of the typical phenomena as such is known as *inquisitive contractions* (Olsson & Westlund, 2006). To keep it straightforward, we will elaborate with the following single-agent scenario, and show how IPM can be used to model this dynamic process.

Scenario: Alice believes there is a seminar this afternoon. She is wondering if Bill is coming. Suddenly Charlie pops up and says to her that there won't be a seminar today. Alice doesn't totally buy it, yet she also starts wondering if there will be a seminar this afternoon.

Analysis: The inquisitive state of Alice was the issue $?S \wedge ?Cb$. However, since she believed that there is a seminar (S), she was actually considering whether Bill is coming at first, as shown in Fig.1(a). After being informed by Charlie, her belief is weakened – she thinks it is quite possible that there won't be a seminar ($[\downarrow S]$), hence her belief state fully covers her epistemic state, and she goes back to entertain the issue $?S \wedge ?Cb$, as in Fig.1(b).



(a) $?b \wedge ?c$ (b) $?c$

Fig. 1. Contraction

Therefore, IPM can be used to model inquisitive contractions, and in general, the dynamics of belief changes interacting with the agents' research agenda. However, there is more to be done to increase the descriptive power. First let us consider two motivating scenarios.

1.4 Motivating Scenarios: First Things First!

Scenario 1 : Alice asked Bill and Charlie to have dinner in a restaurant. Alice doesn't know whether they will come, and she wants to know. Moreover, Alice knows Bill likes Pasta, and Charlie likes rice. She knows that both prices of rice and pasta are 7 or 8 euros, but she is not sure which is which, and she wants to know. Then Bill called, telling her that he is coming, but Charlie is not. Now Alice feels more *urgent* to know the price of pasta.

Scenario 2 : Alice believes that there is a seminar this afternoon, and she is wondering whether Bill is coming. Now Charlie pops up again and says to her that there won't be a seminar today. Alice doesn't totally buy it, yet now she wants to know whether there is a seminar *first*.

Here and henceforth we will refer to Scenario 1 as the "dinner scenario" and Scenario 2 as "seminar scenario". As is shown in both cases, a formalization of the notion of *urgency* of questions is in need to capture the change of research agendas as described. In the next section, we will introduce the *Inquisitive Plausibility-Urgency Model*, which is a modest extention to IPM, and will thus provide us with more dynamics for belief revision and questions.

2 Inquisitive Plausibility-Urgency Model (IPUM)

2.1 The *Urgency* Relation

As mentioned in the final part of last section, our goal here is to (qualitatively) model the degree of *urgency* of different questions. How can we achieve it? Recall that in Epistemic Plausibility Model, the meaning of a proposition is characterized as the set of possible worlds on which the proposition is true, and we capture the agents' *beliefs* by defining a plausibility relation between possible worlds, which is based on an agent's attitudes toward different worlds. In inquisitive semantics, the meaning of a question is modeled as the set of information states, therefore, in accordance, we should be able to model an agent's attitude towards a question as a collective manifestation of her attitude towards its resolutions. Moreover, in practice, one is more *urgent* to know the answer of a question means he/she will feel more satisfied/relieved after hearing it. Therefore, just like the plausibility relation, we can define an *urgency* relation $\preccurlyeq_a$ between information states in a *conditional* point of view, as follows. Fixing any two information states s, t:

$s \preccurlyeq_a t \Longleftrightarrow a$ is at least as urgent/satisfied/relieved to know t as she is for s

By adding this urgency relation to IPM, we will get what we want – the *Inquisitive Plausibility-Urgency Model*.

2.2 The Model

Definition 5. *An inquisitive plausibility-urgency model* $\mathcal{M}$ for a set of agents $\mathcal{A}$ is a tuple

$$\mathcal{M} = (W, \{\sigma_a\}_{a\in\mathcal{A}}, \{\Sigma_a\}_{a\in\mathcal{A}}, \{\leq_a\}_{a\in\mathcal{A}}, \{\preccurlyeq_a\}_{a\in\mathcal{A}}, \|\cdot\|)$$

where

- W is a set of possible worlds.
- an epistemic map σ_a for each agent.
- an inquisitive map Σ_a for each agent.
- a plausibility map $\leq_a \subseteq W \times W$ for each agent.
- an urgency map $\preccurlyeq_a \subseteq \mathcal{P}(W) \times \mathcal{P}(W)$ over $\Sigma_a(w)$ for each agent a at w, which is a converse-well-founded total preorder on information states. Also, for each $s, t \in \mathcal{P}(W)$ *s.t.* $s \preccurlyeq_a t$, if $t' \subseteq t$, then $s \preccurlyeq_a t'$.
- a valuation function $\|\cdot\|$.

Similar to Epistemic Plausibility Model as in Def.1, we can define modalities of *urgent-entertaining* (UE_a), *strong urgent-entertaining* (Sue_a) and *conditional urgent-entertaining* UE_a^Q. Given some issue $\mu :=?\{\alpha_1, ..., \alpha_n\}$:

- $\mathcal{M}, w \vDash UE_a\mu \Longleftrightarrow Max_{\preccurlyeq_a}\{t \in \mathcal{P}(W) : t \in \Sigma_a(w)\} \subseteq \mu$
- $\mathcal{M}, w \vDash Sue_a\mu \Longleftrightarrow$ for any $s, t \in \mathcal{P}(W)$, if $s \notin \mu$ and $t \in \mu$, then $s \prec_a t$.

- $\mathcal{M}, w \vDash UE_a^Q\mu \iff$ for any $t \in Max_{\preccurlyeq_a}\{t \in \mathcal{P}(W) : t \in \Sigma_a(w)\}$ and $t \subseteq \| Q \|$, $t \in \mu$
- $\mathcal{M}, w \vDash UE_a^B\mu \iff$ for any $t \in Max_{\preccurlyeq_a}\{t \in \mathcal{P}(W) : t \in \Sigma_a(w)\}$ and $t \subseteq bel_a(w)$, $t \in \mu$

Also similar to E_a^B, UE_a^B is the urgent-entertaining-over-belief modality, which is used to address issues the agent is urgently entertaining over the current belief state.

There are a few notions worth mentioning. In constructing a model regarding specific situations, we take complete resolutions to an issue that is urgently entertained as equivalently satisfying, i.e. given $\mu = ?\{\alpha_1, ..., \alpha_n\} \in \Sigma_a(w)$ and $\mathcal{M}, w \vDash UE_a\mu$, for any $s, t \in \mu$, there is $s \preccurlyeq_a t$ and $t \preccurlyeq_a s$, or rather, $s \approx_a t$, where $\approx_a$ is the equivalence relation in terms of the urgency order. Also, as $\leq_a$ in Epistemic Plausibility Model can fully describe the Epistemic map σ_a, the urgency relation $\preccurlyeq_a$ also determines the inquisitive state Σ_a, in the sense that the degree of urgency indicates certain inquiries. Formally, it can be achieved by simply intersecting all the downward closures of the equivalent sets which don't support each other. Last but not least, potential changes of research agendas, or in a sense the inquiry strategies are pre-encoded in the model. In particular, differences in urgency may be revealed at different level of information, which result in a more fine-grained reaction of an agent towards different information pieces. This feature can lead us to a solution to the dinner scenario, which requires different reactions of Alice given different informations (whether Charlie or Bill is coming). We will show how IPUM provides us with a relatively elegant formalization of both scenarios mentioned in §1.4 in the next section. Before that, we will first introduce a bonus effect coming with IPUM.

3 Towards More Dynamics of Questions

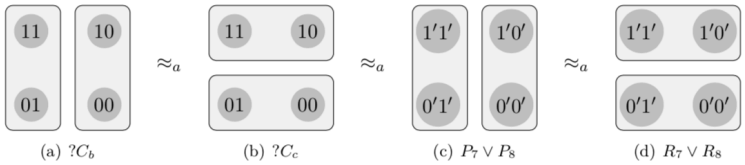
In this section, we will go through some applications of IPUM, by resolving the motivation scenarios introduced in §1.4. By resolving the dinner issue, we demonstrate that the urgency order among different questions can be revealed on partial resolutions to the total issue, therefore capture the reaction of an agent toward certain pieces of information. The solution to the seminar scenario will present a more fine-grained formalization of inquisitive contraction. Starting from these applications, it is straightforward to extend to more dynamics of questions interacting with knowledge/belief change using IPUM.

Scenario 1: Alice asked Bill and Charlie to have dinner in a restaurant. Alice doesn't know whether they will come, and she wants to know. Moreover, Alice knows Bill likes Pasta, and Charlie likes rice. She knows that both prices of rice and pasta are 7 or 8 euros, but she is not sure which is which, and she wants to know. Then Bill called, telling her that he is coming, but Charlie is not. Now Alice feels more *urgent* to know the price of pasta.

Analysis 1: Initially, Alice was wondering whether Bill and Charlie was coming, as well as the price of rice and pasta. She had neither belief nor urgency on either of the two issues, therefore her research agenda is the total issue as pictured in Fig.2(a). Since Alice knows that Bill likes Pasta and Charlie likes rice, the urgency relation between resolutions to the price issues should vary according to resolutions to the coming issue. For instance, under the condition that Bill is coming but Charlie is not ($Cb \wedge \neg Cc$), Alice should be more eager to know the price of pasta. Therefore, we can construct a strict urgency order between resolutions to the price issue among information states that are already resolutions to the coming issue, as shown in Fig.3(a). After Bill called, Alice came to believe that Bill was coming but not Charlie, therefore Fig.3(a) pictures just the research agenda of Alice at that moment. Based on this model, the following propositions are true:

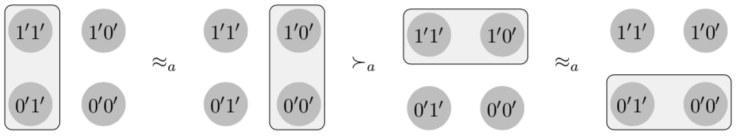
- $\mathcal{M}, w \vDash E_a(?C_b \wedge ?C_c \wedge (P_7 \vee P_8) \wedge (R_7 \vee R_8))$
- $\mathcal{M}, w \vDash UE_a^{C_b \wedge \neg C_c}(P_7 \vee P_8)$
- $\mathcal{M}, w \vDash Sue_a^{C_b \wedge \neg C_c}(P_7 \vee P_8)$

That is, Alice has an inquisitive state consisting of both coming issues and price issues; conditioning on the information that Bill is coming but Charlie is not, Alice will urgently entertain the price of pasta; in fact, she will "strongly urgently" entertain this issue.



(a) $?C_b$ (b) $?C_c$ (c) $P_7 \vee P_8$ (d) $R_7 \vee R_8$

Fig. 2. Dinner Scenario - before



(a) P_7 (b) P_8 (c) R_7 (d) R_8

Fig. 3. Dinner Scenario - after

Scenario 2: Alice believes that there is a seminar this afternoon, and she is wondering whether Bill is coming. Now Charlie pops up again and says to her that there won't be a seminar today. Alice doesn't totally buy it, yet

now she wants to know whether there is a seminar *first*.

Analysis 2: Alice's epistemic goal is to know whether there is a seminar today and whether Bill is coming. Moreover, Alice thinks the former as the more urgent issue. Therefore her inquisitive state along with the urgency order can be shown as in Fig.4(a). However, initially Alice believed that there is a seminar, which makes the latter the only issue she is entertaining (see also Fig.1.3). After withdrawing this belief, the urgency order reveals itself. Hence we have the following results:

- $\mathcal{M}, w \vDash E_a^B(?Cb)$
- $\mathcal{M}, w \vDash [\downarrow S]UE_a(?S)$

That is, based on her belief, Alice is entertaining whether Bill is coming, but when the belief is retracted, she begins to urgently entertain whether there is a seminar.

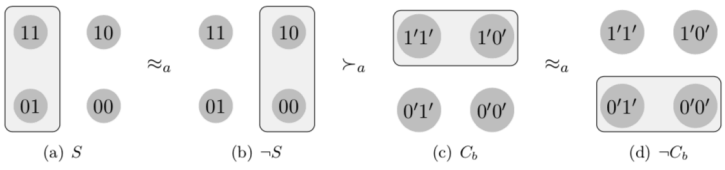


Fig. 4. Seminar Scenario

4 Conclusion

In this paper, we proposed an extension of inquisitive plausibility model named inquisitive plausibility-urgency model by introducing an *urgency* relation between information states. The relation is defined in a conditional manner analogous to the plausibility relation. Using this model, we can capture more dynamics of questions in belief change, such as the reaction of an agent towards (partial) resolutions to the issue she's entertaining, as well as a more fine-grained notion of inquisitive contraction.

The description of IPUM in the paper is rather sketchy, and we hope to further specify the details in a full paper. Meanwhile, based on this basic model, some immediate future direction should be concerned. For one thing, a sound and complete axiomatization (logic) is open for investigation. For another, here we restrict our model in a single-agent setting; in order to extend it to a multi-agent setting, we need to consider additional complications such as the attitudes of an agent towards the issues raised or entertained by other agents, which will hopefully lead us to a more complete picture of insecure communication. We hope this paper can serve as a modest spur that induces revelations of inquisitive potential, as well as its influence on dynamic epistemic/doxastic logic.

Acknowledgements

The authors would like to thank Dr. Alexandru Baltag for the course *Dynamic Epistemic Logic* at the University of Amsterdam. We are also grateful to Dr. Floris Roelofsen, the Inquisitive Semantics group at the ILLC, and the anonymous reviewers for their valuable feedbacks. Last but not least, many thanks to ESSLLI 2018 Grant Committee and EACL for the student grants.

Reference

Baltag, A, R Boddy & S Smets. 2016. Group knowledge in interrogative epistemology. *Outstanding Contributions to Logic: J. Hintikka, Springer* .

Baltag, Alexandru & Sonja Smets. 2006a. Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science* 165. 5–21.

Baltag, Alexandru & Sonja Smets. 2006b. The logic of conditional doxastic actions: A theory of dynamic multi-agent belief revision. In *Proceedings of esslli workshop on rationality and knowledge*, 13–30.

Van Benthem, J. & Ş. Minică. 2009. Toward a dynamic logic of questions. In X. He, J. Horty & E. Pacuit (eds.), *Logic, rationality, and interaction*, 27–41. Springer.

Ciardelli, Ivano. 2009. Inquisitive semantics and intermediate logics. Master Thesis, University of Amsterdam.

Ciardelli, Ivano, Jeroen Groenendijk & Floris Roelofsen. 2013. Inquisitive semantics: a new notion of meaning. *Language and Linguistics Compass* 7(9). 459–476.

Ciardelli, Ivano & Floris Roelofsen. 2014. Issues in epistemic change. Talk at *European Epistemology Network Meeting*, Autonomous University of Madrid, July 2014.

Enqvist, Sebastian. 2010. Contraction in interrogative belief revision. *Erkenntnis* 72(3). 315–335.

Groenendijk, Jeroen & Floris Roelofsen. 2009. Inquisitive semantics and pragmatics. In Jesus M. Larrazabal & Larraitz Zubeldia (eds.), *Meaning, content, and argument: Proceedings of the ILCLI international workshop on semantics, pragmatics, and rhetoric*, www.illc.uva.nl/inquisitive-semantics.

Hintikka, J. 2007. *Socratic epistemology: explorations of knowledge-seeking by questioning*. Cambridge University Press.

Olsson, Erik J & David Westlund. 2006. On the role of the research agenda in epistemic change. *Erkenntnis* 65(2). 165–183.

Schaffer, Jonathan. 2005. Contrastive knowledge. *Oxford studies in epistemology* 1. 235–271.

Van Benthem, Johan. 2007. Dynamic logic for belief revision. *Journal of applied non-classical logics* 17(2). 129–155.

Van Ditmarsch, Hans, Wiebe van Der Hoek & Barteld Kooi. 2007. *Dynamic epistemic logic*, vol. 337. Springer Science & Business Media.

Definiteness in Shan

Mary Moroney

Cornell University, Ithaca, New York, USA
mrm366@cornell.edu
http://conf.ling.cornell.edu/mmoroney/about.html

Abstract. Shan, a Southwestern Tai language spoken in Myanmar, Thailand, and nearby countries, uses bare nouns to express both unique and anaphoric definiteness, as identified by [11]. This novel data pattern from the author's fieldwork can be analyzed by adding an anaphoric type shifter, ι^x, to the available type shifting operations defined by [2] and [3]. It also demonstrates that the consistency test is not sufficient to determine what counts as a definite determiner for a language.

Keywords: definiteness · type-shifting · Tai language.

1 Introduction

[11] proposes that there are two types of definiteness expressed by German, corresponding to the contracted (*weak*) and non-contracted (*strong*) preposition + definite article combinations—e.g., *vom* ('by the', *weak*) and *von dem* ('by the', *strong*). In (1), the speaker and listener know that there is only one mayor in the context. Since the mayor is unique in the context, the weak definite article form, *vom* ('by the') is used and the strong form is infelicitous.

(1) Weak versus strong articles in German ([11]: (42))

Der Empfang wurde **vom** / **#von dem** Bürgermeister
the reception was by-the$_{weak}$ / by the$_{strong}$ mayor
eröffnet.
opened

'The reception was opened by the mayor.'

[11] claims that the split between strong and weak definite forms fits well with the types of definiteness described by [5], which grouped definiteness into four categories: immediate situation (current non-linguistic context), larger situation (broader non-linguistic context), anaphoric/familiar, and bridging (associative anaphora). [11] says that when a noun is unique in an immediate situation or larger situation context, German uses the weak form of the definite article, and in anaphoric contexts it uses the strong form. For the bridging category, he discusses two types: producer-product and part-whole bridging, which I will call

'product-producer' and 'whole-part' bridging, respectively.[1] The strong form is used in product-producer situations and the weak form is used in whole-part bridging. In addition to the categories discussed by [5], [11] adds that donkey anaphora uses the strong form of the definite article. Table 1 gives examples of these categories and the article form used for German. These will be discussed more in the following section.

Table 1. Types of definiteness described by [11], citing [5]

Type of Definite Use	Example	German
Unique in immediate situation	the desk (uttered in a room with exactly one desk)	weak
Unique in larger situation	the prime minister (uttered in the UK)	weak
Anaphoric	John bought a book and a magazine. The book was expensive.	strong
Bridging: Product-producer	John bought a book today. The author is French.	strong
Bridging: Whole-part	John was driving down the street. The steering wheel was cold.	weak
Donkey anaphora	Every farmer who owns a donkey hits the donkey	strong

2 Uniqueness and Anaphoricity

[11] claims that the weak definite article in German expresses *uniqueness*. This can be uniqueness in an immediate situation, as in (2), or in a larger or global context, described further below. In (2), there is only one glass cabinet in the immediate context, so the weak definite must be used.

(2) German: Unique in immediate situation ([11]: (40))

Das Buch, das du suchst, steht **im** / **#in dem**
the book that you look-for stands in-the$_{weak}$ / in the$_{strong}$
Glasschrank.
glass-cabinet

'The book that you are looking for is in the glass-cabinet.'

The strong definite article expresses *familiarity/anaphoricity*. In (3), the first sentence introduces a writer and a politician into the discourse context. In the second sentence *von dem Politiker* ('from the politician') is used to refer back to the politician. The strong definite form must be used in this context.

[1] A review noted that what [11] calls 'part-whole' bridging would more correctly be called 'whole-part' bridging, and agreeing with their assessment, I will use that and 'product-producer' instead of 'producer-product' for the same reasons.

(3) German: Anaphora ([11]: (23))
Hans hat einen Schriftsteller und **einen Politiker** interviewt. Er hat
Hans has a writer and a politician interviewed He has
#vom / **von dem** **Politiker** keine interessanten Antworten
from-the$_{weak}$ / from the$_{strong}$ politician no interesting answers
bekommen.
gotten
'Hans interviewed a writer and a politician. He didn't get any interesting answers from the politician.'

Looking at Mandarin and Thai, [7] and [8] show that these languages use bare nouns in the same places where German would use the weak definite article, and noun phrases modified by a classifier and demonstrative where German would use the strong definite article. Examples (4) and (5) show the use of the bare noun in a unique situation in Mandarin and Thai, respectively.

(4) Mandarin: Unique in immediate sit. ([7]: (12b), citing [1]: 510)
Gou yao guo malu.
dog want cross road
'The dog(s) want to cross the road.'

(5) Thai: Unique in immediate situation ([8]: (2))
mǎa kamlaŋ hàw.
dog prog bark
'The dog is barking.'

In (6) and (7), are the Mandarin and Thai examples using demonstratives to express familiarity/anaphoricity. In (6a), a boy and a girl are introduced into the discourse context. (6b) and (6c) use *na ge nasheng* ('the/that boy'), a noun modified by a classifier and demonstrative, to refer back to the boy. In Mandarin there is a contrast between the subject and object position. The classifier and demonstrative are optional in subject position, but not in object position, as shown in (6b) and (6c). [7] claims that this is because the Mandarin subject is a topic, which negates the need for an antecedent index.

(6) Mandarin: Narrative sequence (Anaphoric) ([7]: (16a,b,d))
a. jiaoshi li zuo-zhe **yi ge nansheng** he **yi ge**
classroom inside sit-prog one clf boy and one clf
nüsheng,
girl
'There is a boy and a girl sitting in the classroom...'
b. Wo zuotian yudao #(**na ge**) **nansheng**
I yesterday meet that clf boy
'I met the boy yesterday.'
c. (**na ge**) **nansheng** kan-qi-lai you er-shi sui zuoyou.
that clf boy look have two-ten year or-so
'The boy looks twenty-years-old or so.'

In the Thai example in (7), (7a) introduces a student into the discourse context. In (7a), *nákrian khon nán* ('that boy') is used to refer to the boy. (7b) suggests that the demonstrative is required even in subject position for Thai.

(7) Thai: Narrative sequence (Anaphoric) ([8]: (17))
mî̂awaan phŏm cəə kàp **nákrian khon niŋ**.
yesterday 1st meet with student clf indef
'Yesterday I met a student'

a. (**nákrian**) **khon nán** / (**kháw**) chalàat mâak.
student clf that / 3p clever very
'That student/(s)he was very clever.'

b. #**nákrian** chalàat mâak.
student clever very
'Student are very clever.'

2.1 Associative Anaphora (Bridging)

[11] shows that in German, there is a split between whole-part and product-producer bridging in terms of definiteness marking: whole-part bridging uses the weak definite and product-producer bridging uses the strong definite. [7] and [8] show that Mandarin and Thai patterns with German, using the bare noun in whole-part examples (weak definiteness) and the demonstrative in product-producer examples (strong definiteness). In this section and the following one, only the Thai data is shown to conserve space. In (8), *thábian* ('sticker') cannot be modified by a demonstrative. This parallels the use of the weak definite for whole-part bridging in German.

(8) Thai: Whole-part bridging ([8]: (11))
rót khan nán thùuk tamrùat sàkàt phrɔ́ʔ mâj.dâj tìt
car clf that adv.pas police intercept because neg attach
satikəə wáj thîi **thábian** (**#baj nán**).
sticker keep at license clf that
'The car was stopped by police because there was no sticker on the license.'

In (9), the producer *náktɛ̀ɛŋklɔɔn* ('poet') must be modified by a demonstrative. This parallels German's use of the strong definite for product-producer bridging.

(9) Thai: Product-producer bridging ([8]: (12))
ʔɔɔl khít wâa **klɔɔn** bòt nán prɔ́ʔ mâak, mɛ̂ɛ-wâa kháw cà
Paul thinks comp poem clf that melodious very although 3p irr
mâj chɔ̂ɔp **náktɛ̀ɛŋklɔɔn** #(**khon nán**).
neg like poet clf that
'Paul thinks that poem is beautiful, though he doesn't really like the poet.'

2.2 Donkey anaphora

In cases of donkey anaphora, [11] claims that German uses the strong article to refer to nouns introduced in the first part of the construction. Similarly, in Thai and Mandarin, a demonstrative is required in those positions ([7]; [8]). For the Thai example in (10), using the bare noun to refer back to the buffalo gives the sentence a generic meaning 'Every farmer that has a buffalo hits buffalo'.

(10) Thai: Donkey anaphora ([8]: (23))

chaawnaa thúk khon thîi mii **khwaai tua nɨŋ** tii **khwaai tua**
farmer every CLF that have buffalo CLF INDEF hit buffalo CLF
nán
that

'Every farmer that has a buffalo hits it.'

Table 2 summarizes the patterns of definiteness expression in German, Thai, and Mandarin. Examples of all the contexts described by [11] cannot be included due to space limitations, but they can be found in the cited sources.

Table 2. Expressions of definiteness in German, Thai, and Mandarin

Type of Definite Use	German ([11])	Thai ([8])	Mandarin ([7])
Immediate situation	weak	bare	bare
Larger situation	weak	bare	bare
Anaphoric	strong	dem.	dem.
Bridging: Product-producer	strong	dem.	dem.
Bridging: Whole-part	weak	bare	bare
Donkey anaphora	strong	dem.	dem.

3 Shan

Like Mandarin and Thai, Shan, a Southwestern Tai language spoken in Myanmar, uses the bare noun in unique situations, as shown in (11) and (12).[2,3] In (11), there is a single teacher in the context, so it must be referred to using a bare noun. In (12), world knowledge tells us that there is only one sun, so a bare noun is used to refer to the sun. The demonstrative is not felicitous in either case.

[2] Data for this paper comes from the author's fieldwork in Chiang Mai, Thailand from January 2018 to present, working with a speaker from Keng Tawng City in Shan State, Myanmar, who has lived in Thailand for over 10 years. Data was collected using a variety of elicitation methods: story translation, stories based on storyboards, felicity judgments on grammatical sentences in specific contexts.

[3] Glossing conventions: 1: first person, 3: third person, CL: classifier, COMP: complementizer, IMPF: imperfect, NEG: negation, SG: singular

(11) Shan: Unique in immediate situation
(Context: classroom with just one teacher)
Náaŋ Lɤ̌n ʔàm tsaaŋ kwàa hăa **khúsɔ̃n** (**#kô nân**)
Ms. Lun neg able go find teacher cl.person that
'Ms. Lun cannot find the teacher.'

(12) Shan: Unique in larger situation
kăaŋwán (**#hòj nân**) lŏŋ hɤ sɔ̀ŋ.
sun cl.round that very bright glitter
'The sun is very bright.'
(Speaker comment on the demonstrative: there is more than one sun)

3.1 Anaphora

Unlike Mandarin and Thai, Shan can use the bare noun in anaphoric contexts such as a narrative sequence. In (13), the first sentence introduces a man into the discourse context. In following sentences, the man can be referred back to either using a bare noun, *phu-tsáaj* ('man'), or using a bare noun modified by a classifier and demonstrative, *phu-tsáaj kô nân* ('that man').[4]

(13) Shan: Narrative Sequence (Anaphora)
phu-tsáaj kô nɯŋ kwàa ti hâan khăaj măa tàa sûɯ maă
person-man cl.person one go at store sell dog for buy dog
ʔɔ̀n tŏ nɯŋ păn luk jíŋ mán-tsáaj... **phu-tsáaj**
small cl.animal one give child girl 3-man person-man
(**kô nân**) khún tɔ̀p waa,
cl.person that back respond that
'A man went to a dog store to buy a puppy for his daughter... The/that man replied,'

In (14), the first sentence introduces a notebook and cup of water into the discourse context. The second sentence refers back to each of them using a bare noun. Here the anaphoric nouns are in object position, but this position does not require that a demonstrative be used. In this way, Shan is different from Mandarin or Thai. The demonstrative is allowed, but it sounds awkward to use a demonstrative for both the water cup and notebook in the second sentence.

[4] A reviewer noted that the examples from Mandarin and Thai are not equivalent in that the Mandarin one introduces two individuals apart from the speaker, and the Thai one only introduces one other individual. For Shan, I have included both types of examples. In (13) there is only one individual, and in (14) there are two.

(14) Shan: Narrative Sequence (Anaphora)
pâp măaj lɛ **kɔ́k nâm** jù wâj nɤ̌ phɤ̌n. khaa qăw **kɔ́k nâm**
book note and cup water IMPF stay on desk 1.SG take cup water
(**nân**) he sàj/saù **pâp** (**nân**).
spill in book CL.BOOK that

'There is a notebook and a cup of water on the desk. I spilled the/that cup of water onto the/that notebook.'

3.2 Bridging

Mandarin, Thai, and German use the weak/bare form of the nominal in whole-part bridging and the strong/demonstrative form in product-producer bridging. Shan, instead, does not use different nominal expressions in whole-part bridging versus product-producer bridging. A bare noun can be used in both situations. (15) shows that a bare noun is possible for whole-part bridging in Shan.

(15) Shan: Whole-part bridging
khúsvɔn kwàa tsú **hɤ́ntɤk** lăŋ nân sĕ tɔ̀j **pháktŭ**
teacher go to building CL.BUILDING that and knocked door
hɔ̂ŋ tsaw hɤ́n
call owner building

'The teacher approaches that building and knocked on the door to call the owner.'

Whole-part bridging constructions in Shan often have the 'whole' as part of the word for the 'part'. It is not always clear whether it simply anaphoric with the 'whole' possessing the part or involves bridging to a real noun compound. (16) shows an example of this where *naasɤ̂ pâplik* ('book cover') contains the word *pâplik* ('book'). While it is possible to modify the noun with a demonstrative, the demonstrative is referring to the book rather than the cover. It does not seem possible to modify the bridged noun with a demonstrative.

(16) Shan: Whole-part bridging
mɛ́w wĕnkjɔ́k sàj nɤ̌ **pâplik** ʔăn mí nɤ̌ phɤ̌n măa kwàa tsɔ́m
cat jump in on book COMP exist on table dog go follow
thɛŋ. thŭŋ ti hét haj **naasɤ̂ pâplik** (**nân**) kokòmkolĭn kwàa
again until COMP do cause cover book that dirty go
seŋ
completely

'The cat jumped onto the book that was on the table. The dog followed again which made the book cover/cover of that book completely dirty.'

(17) shows that a demonstrative is not necessary for product-producer bridging either. The 'producer', *kɔ́ntɛmlik* ('author') can be bare or modified by the demonstrative, *kɔ̂ nân*. From the classifier we can tell that this demonstrative modifies 'author' not 'book'.

(17) Shan: Product-producer bridging
mʏwáa khú ʔàan **pâplik pʊ́n táj**. khúsɔ̌n pěn ʔójkɔ̂ kăn
yesterday teacher read book history Tai teacher be friend together
táŋ **kóntɛmlik** (**kɔ̂ nân**)
with author CL.PERSON that

'Yesterday, the teacher read a Tai (Shan) history book. The teacher is friends with the/that author.'

3.3 Donkey anaphora

German, Thai, and Mandarin use the strong/demonstrative form of the nominal to refer anaphorically to a nominal in donkey anaphora. Unlike the other languages, Shan does not use a demonstrative or strong definite article in this situation. In (18), when 'cat' (*mέw*) is referred to anaphorically, a bare noun is used. It is not felicitous to modify it with a demonstrative because that forces a singular reading, which sounds awkward in this sort of generic sentence.

(18) Shan: Donkey anaphora
măa ku tŏ nâj pɔ́ hăn **mέw** năj tĕ lɯp lám
dog every CL.ANIMAL this if/when see cat then will follow chase
mέw (***tŏ nân**) tàasè
cat CL.ANIMAL that always

'Every dog, if it sees a cat will always chase the cat.'

If we wanted to use a demonstrative in this sort of example, a structure like (19) would be possible, but, again, the classifier-demonstrative modification is not necessary. The difference between these two examples is that in (18) it is dogs being quantified over, leaving 'cat' as unspecified for plurality and thus awkward with a singular anaphor. In (19), *tŏ lăj* ('which one') quantifies over individual cats making it compatible with a singular anaphor.

(19) Shan: Donkey anaphora
măa nâj hăn **mέw tŏ lăj** kɔ tĕ lɯp **mέw**
dog this see cat CL.ANIMAL which even will follow cat
(**tŏ nân**) tàasè
CL.ANIMAL that always

'Dogs, whichever cat they see they will always chase the/that cat'

Table 3 summarizes the different expressions of definiteness found in German, Thai, Mandarin, and Shan. This section has investigated the pattern of definiteness found in Shan in specific contexts that have shown different patterns of expression across languages. Shan allows for the bare noun to be used in all of the contexts described by [11]. Even contexts like anaphora and product-producer bridging allow for bare nouns where Thai and Mandarin do not. For contexts were the noun is unique in a situation or with whole-part bridging, a demonstrative cannot modify the noun, just like in Thai and Mandarin.

Table 3. Expressions of definiteness in German, Thai, Mandarin, and Shan

Type of Definite Use	German ([11])	Thai ([8])	Mandarin ([7])	Shan
Immediate situation	weak	bare	bare	bare (11)
Larger situation	weak	bare	bare	bare (12)
Anaphoric	strong	dem.	dem.	**bare** (13-14)
Bridging: Product-producer	strong	dem.	dem.	**bare** (17)
Bridging: Whole-part	weak	bare	bare	bare (16)
Donkey anaphora	strong	dem.	dem.	**bare** (18-19)

4 Analysis

Following [2] and [3], [4] summarizes the available interpretations of bare nouns in languages without articles, claiming they can have a kind reading, a narrow scope existential reading, and a definite reading. This appears to be consistent with what is found in Shan. [2] claims that bare nouns in article-less languages without number marking, like Shan, obligatorily have an e-type, kind denotation. However, [4] allows for these mass nouns to undergo type shifting using $^{\cup}$ so they can then type-shift using ι to get a definite reading separate from the kind reading. For now, I will assume this, following [4], but this topic should be considered in future work. The type-shifting operators described by [2] and [3], $^{\cap}$, ι, and $\exists$, are defined below:

(20) Type shifting operators ([3]): $\langle e,t\rangle \rightarrow e/\langle\langle e,t\rangle,t\rangle$
   a. $^{\cap}$: $\lambda P \lambda s \iota x[P_s(x)]$
   b. ι: $\lambda P \iota x[P_s(x)]$
   c. $\exists$: $\lambda P \lambda Q \exists x[P_s(x) \wedge Q(x)]$

[3], revising [2], proposes that the type shifting operators follow a hierarchy, where kind-forming $^{\cap}$ and entity forming ι must be ruled out before $\exists$ becomes available, this is described in (21). The justification is that using $^{\cap}$ or ι is a less drastic change because it does not introduce quantificational force. [3] claims that bare nouns are equally allowed to form kinds or entities, so they must be ranked equally. [3] and [2] use the Blocking Principle, defined in (22), to identify what type shifting is available in what language. If a language has an overt determiner form of a type shifter—e.g., *the* in English is said to correspond to ι—then covert type shifting using that operator is unavailable.[5]

[5] In a language where there are are no determiners, you would expect all type shifting operations to be available, but according to [3], $\exists$-type shifting does not occur in these languages because of the ranking described in Meaning Preservation below. The existential interpretation comes form Derived Kind Predication. This is an interesting subject for future investigation, but not addressed here.

(21) Meaning Preservation: $\{\cap, \iota\} > \exists$

(22) Blocking Principle [3]: For any type shifting operation π and any X: $*\pi(X)$ if there is a determiner D such that for any set X in its domain, $D(X) = \pi(X)$.

[7] follows [11] in claiming the existence of two types of definiteness. In trying to account for the obligatory use of the demonstrative in some definite environments in Mandarin, [7] defines the unique and anaphoric definites as in (23), where (23a) is the type shifting operation ι and (23b) is the denotation of the demonstrative in Mandarin.[6] [8] claims that since English expresses both unique and anaphoric definites using *the*, *the* is ambiguous for the unique and anaphoric definite meaning.

(23) a. Unique definite article:
$[\![\iota]\!] = \lambda s_r.\lambda P_{\langle e,\langle s,t\rangle\rangle}. : \exists!x[P(x)(s_r)].\iota x P(x)(s_r)$
b. Anaphoric definite article: ι^x
$[\![\iota^x]\!] = \lambda s_r.\lambda P_{\langle e,\langle s,t\rangle\rangle}.\lambda Q_{\langle e,t\rangle}. : \exists!x[P(x)(s_r) \wedge Q(x)].\iota x P(x)(s_r)$

It is clear from the data that the Shan demonstrative does not fill the roll of anaphoric definite determiner since it is not obligatory in all anaphoric contexts as in Thai. I propose, instead that Shan has a null anaphoric type shifter ι^x in addition to the ι type shifter.

This analysis raises the question: Why does the Shan demonstrative not count as a determiner for the purposes of the Blocking Principle, but the Thai and Mandarin ones do? We might expect the Shan demonstrative to pattern differently from the Mandarin and Thai demonstratives in terms of the Consistency test. [3] uses the Consistency test from [9] to distinguish between demonstratives and true definites. For demonstratives, you can introduce two of the same NPs modified by the demonstrative with contradictory predicates, and there is no contradiction. For definite determiners, doing this would create a contradiction. According to this test, Shan has a demonstrative, not a definite determiner, as shown in (24). However, the Thai demonstrative also passes this test, as in (25).[7]

(24) Shan: Consistency test
(Context: I am holding a white cup and a black cup.)
kɔ́k hòj **nâj** pěn sǐ khăaw. **kɔ́k hòj** **nâj** pěn sǐ
cup CL.ROUND this be color white cup CL.ROUND this be color
lăm.
black
'This cup is white. This cup is black.'

[6] This definition differs from [11] in that the index is defined as a property rather than an individual, but I will not be concerned with this distinction for this analysis.

[7] Mandarin passes the consistency test too, but the data is not included here to conserve space.

(25) Thai: Consistency test ([8], citing [10])

dèk khon nán nɔɔn yùu tɛ̀ɛ **dèk khon nán** mâi.dâi nɔɔn yùu.
child CLF that sleep IMPF but child CLF that NEG sleep IMPF

'That child is sleeping but that child is not sleeping.' (cf. #the)

According to a native Thai speaker, (25) sounds contradictory out of the blue, but fine with deixis. This test does not seem sufficient to distinguish between what counts as a definite for the Blocking principle. This is not that surprising since the consistency test relies on deixis which is not something that comes into play in anaphoric uses of demonstratives.

I would argue that the Shan bare noun/demonstrative contrast parallels the English *the*/demonstrative contrast. The difference comes from the fact that the bare noun in Shan can denote a broader range of things, which might lead to more disambiguation using the demonstrative. We would then expect the use of the demonstrative in Shan to convey some special meaning beyond ι in the same way the English demonstrative can.

5 Conclusion

Shan can use a bare noun to express both unique and anaphoric definiteness. In fact, the bare noun in Shan behaves much like the English article *the*. Though languages like Thai and Mandarin are similar to Shan in lacking overt definite articles and plural morphology, Shan does not pattern together with these two languages in that its demonstrative does not function as the primary marking of anaphoric definiteness. The pattern in Shan is likely to be found in other languages without articles, like Japanese and Russian.

This paper has also shown that the Consistency test does not seem able to distinguish what words count as determiners, so future work should address distinguishing between demonstratives and definite determiners. In Mandarin and Thai, the demonstrative counts as the determiner denoting ι^x, so the demonstrative is obligatory in expressing this meaning. I argue that in Shan the demonstrative does not count as a determiner ι^x, so a bare noun can type shift using ι^x. It seems, then, that the anaphoric definite, ι^x, could be included as one of the available type shifting operations.

This work in conjunction with the work by [11], [8], and [7] brings up the connection between form and meaning. In Mandarin, Thai, and German there seems to be a connection between the obligatory use of a strong determiner/demonstrative and the need for an anaphoric index in the meaning. In Shan, that connection is unidirectional: if there is a demonstrative there must be an anaphoric index, but the lack of a demonstrative does not mean there is no anaphora involved. We might then wonder if we want to say there are two covert ι's in this language. The importance of separating them is apparent in contexts where their meanings are different. In German, it is possible to see an overt contrast between the unique and anaphoric reading, as in (26).

(26) Two definite articles in German ([11]: 268)

Wenn [**ein ausländischer Präsident**]$_1$ [**Barack Obama**]$_2$ im
when a foreign president Barack Obama in-the$_{weak}$
WeißHaus besuch, wird **vom**$_1$ / **von dem**$_2$ **Präsidenten** eine
White House visits by-the$_{weak}$ / by the$_{strong}$ president a
Rede gehalten
speech given

'When a foreign president visits Barack Obama in the White House, the president gives a speech.'

The English translation is also ambiguous, as is the Shan version. This ambiguity must come from differences in the semantic denotations, which could correspond to ι and ι^x. Without some sort of distinction we cannot explain why such examples are ambiguous. The goal of investigating data of this sort is to identify which features to model in order to capture the range of expressions of definiteness across languages. Using these different type shifters is one way we can do this. It opens the question: how many type-shifters do we need to account for definiteness? Future work would be to integrate this analysis into a complete analysis of the interpretation of Shan nouns and further compare with other languages.

6 Acknowledgements

Thanks to Nan San Hwam, Mai Hong, and Sai Loen Kham who provided the Shan data. Thanks also to Sarah Murray, Miloje Despic, the Cornell Semantics Group, and ESSLLI reviews for all their feedback. All errors are my own.

References

1. Cheng, Lisa Lai-Shan and Rint Sybesma: Bare and not-so-bare nouns and the structure of NP. Linguistic Inquiry 30, 509–542 (1999)
2. Chierchia, Gennaro: Reference to kinds across language. Natural language semantics. 6, 339–405 (1998)
3. Dayal, Veneeta: Number marking and (in) definiteness in kind terms. Linguistics and Philosophy 27, 393-450. (2004)
4. Deal, Amy Rose, and Julia Nee: Bare nouns, number, and definiteness in Teotitln del Valle Zapotec. Proceedings of Sinn und Bedeutung. Vol. 21. (2017)
5. Hawkins, John A.: Definiteness and indefiniteness: a study in reference and grammaticality prediction. London, Croom Helm. (1978)
6. Heim, Irene: Artikel und Definitheit. Semantik: Ein internationales Handbuch der zeitgenssischen Forschung, ed. by A. von Stechow and D. Wunderlich, Berlin: de Gruyter. (1991)
7. Jenks, Peter: Articulated definiteness without articles. Linguistic Inquiry 49.501–536. (2018)

8. Jenks, Peter: Two kinds of definites in numeral classifier languages. Semantics and Linguistic Theory (SALT) 25, ed. by Sarah DAntonio, Mary Moroney, and Carol-Rose Little, 103-124, LSA and CLC Publications. (2015)
9. Löbner, Sebastian: Definites. Journal of Semantics 4, 279–326. (1985)
10. Piriyawiboon, Nattaya: Classifiers and Determiner-less Languages: The Case of Thai. University of Toronto, PhD dissertation. (2010)
11. Schwarz, Floraian: Two types of definites in natural language. University of Massachusetts Amherst. (2009)

Compositionality in privative adjectives: extending Dual Content semantics

Joshua Martin

Harvard University

Abstract. Privative adjectives such as *fake* have long posed problems to theories of adjectives in a compositional semantics. In this paper, I argue that a theory like Del Pinal's recent Dual Content semantics, which encodes lexical entries with both an extension-determining component, and a conceptual component, is best equipped to account for privativity while maintaining compositionality. I provide some novel evidence for this system regarding recursive privativity, and some recent experimental data, and introduce an extension to the system to account for pseudo-privative behavior of predicative adjectives.

1 The puzzle of privativity

Kamp (in [1]) argues that all adjectives can be analyzed as functions from properties to properties. While some adjectives can operate as predicates, others do not, and necessarily operate over the extension of the head noun they modify. Thus, to achieve a uniform account, one theory treats them all as the more complex type. As [2] acknowledges, with the advent of new, independently motivated semantic compositional rules beyond function application (e.g. type-shifting or Predicate Modification), this 'generalize to the worst case' approach is no longer necessary, and the simpler adjectives might be assigned a simpler type, with their higher-type behaviors accounted for with more complex compositional operations. The most typical of the 'worst case' categories are privative non-subsective adjectives, which entail the negation of the head noun, such that the intersection of the modified NP and the bare noun is the empty set. The privative behavior of adjectives like *fake* and *stone* in some contexts poses a notable challenge for compositional theories of semantics to grapple with. Since no elements of the bare noun set end up in the modified NP set, it is difficult to evaluate what contribution the bare noun makes to the compositional process, and what kind of operation the privative adjective is performing over it. While the entailment pattern suggests that the adjective is negating the denotation of the noun, it is clearly insufficient to say that a phrase like *fake guns* denotes 'the set of things which are not guns', which would include not only what we understand fake guns to be but also all other non-gun entities in the world.

Partee (see [3]) summarizes this problem, and suggests that what we have heretofore considered privative nonsubsective adjectives are in fact neither of

those things, and are actually a type of subsective adjectives. Their pseudo-privative behavior arises from the fact that they 'coerce' the noun into an expanded meaning to avoid vacuity, and then pick out a subset of that expanded set. She argues that this is necessary to deal with data like (1):

(1) a. A fake gun is not a gun.
b. Is that gun real or fake?

Under most analyses, there is an obvious tension between the acceptability of both of these sentences, such that (1a) can be true while *fake* can also predicated of *that gun* in (1b). Partee argues that this is due to *fake* coercing *gun* into an expanded meaning in (1b), while the unmodified instance of *gun* in (1a) retains is original, limited denotation to the exclusion of fake guns. An expanded denotation of *gun* that can include fake guns is also argued to be necessary for felicitous use of *real*, which would otherwise be vacuous and redundant if all members of the set denoted by *gun* were always real guns.

Her other arguments, based on Polish NP-splitting data, will not be reviewed here, but it is notable that while she motivates coercion, she does not implement it mechanically in any substantive way. How, precisely, is the denotation of a noun coerced into a larger meaning, and perhaps more crucially, what is the larger meaning that a noun like *gun* is coerced into by the presence of *fake*? If it were merely expanded to include the presence of non-gun objects, this would surely overgenerate to a class far larger than what we would naturally consider fake guns. This is the problem that I think the following system addresses more successfully.

2 Introducing Dual Content

Del Pinal's Dual Content semantics (see [4]) preserves a system of composition nearly identical to classical function application, with minimal modifications, by adopting the assumption that the default lexical entries for nouns are notably more complex than in prior systems. On this view, common nouns have a binary semantic structure consisting of their extensional meaning (E-structure) and their conceptual meaning (C-structure). E-structure is the atomic extension-determining component, of the form $\lambda x.\textsc{Stone}(x)$. C-structure does not determine the extension of a noun, but instead consists of 'representations of perceptual features, functional features and genealogical features related to [the noun]' ([4], p. 4). These take the form of a Pustejovsky-style (see [7]) qualia structure. While only E-structure determines the extension of the noun's denotation, C-structure, it is argued, is a necessary component of speaker's linguistic competence in their ability to correctly identify members of a kind and use the term dynamically and productively in different contexts; in a sense, it might be considered an instruction manual for correct and useful application of the linguistic term. A sample entry for *gun* is given below.

(2) $[\![\textbf{gun}]\!] =$
E-structure: $\lambda x.\textsc{Gun}(x)$
C-structure:
$\textsc{Constitutive}$: $\lambda x.\textsc{Parts-Gun}(x)$
$\textsc{Formal}$: $\lambda x.\textsc{Perceptual-Gun}(x)$
$\textsc{Telic}$: $\lambda x.\textsc{Gen}\ e[\textsc{Shooting}(e) \wedge \textsc{Instrument}(e, x)]$
$\textsc{Agentive}$: $\lambda x.\exists e_1[\textsc{Making}(e_1) \wedge \textsc{Goal}(e_1, \textsc{Gen}\ e[\textsc{Shooting}(e) \wedge \textsc{Instrument}(e, x)])]$

Not all of the precise semantics of the C-structure elements in (2) will come into play here; what is necessary is that the C-structure of *gun* encodes that it is composed of gun parts, has the perceptual form of a gun, is generally used in shooting events, and was made with the goal to be used in shooting events. Now that lexical entries are decomposed in this format, we also introduce operations which are able to access specific components of a lexical entry's meaning.

(3) *Qualia functions*: partial functions from the meaning of terms into their respective C-structure denotations, namely, constitutive, formal, telic, and agentive. The qualia functions are Q_C, Q_F, Q_T, Q_A. For example, using the denotation for *gun* in (2): $Q_C([\![\textbf{gun}]\!]) = \lambda x.\textsc{Parts-Gun}(x)$

Adjectives can have a Dual Content structure as well. Intersective adjectives could theoretically be represented in a simpler manner, perhaps with only E-structure, such as $[\![\textbf{red}]\!] = \lambda D_C.\lambda x.D_C(x) \wedge \textsc{Red}(x)$, or even more simply as type $\langle e, t\rangle$ and composing with nouns using Predicate Modification. Privative adjectives, then, make use of these qualia functions, bringing through different elements of the noun's C-structure (either preserved or negated). I will skip over much of Del Pinal's exposition and argument for how he arrives at this eventual lexical entry for *fake*, and simply present the final version. Here, D_C is the domain of 'ordered sets of the E-structure and C-structure of common Ns' ([4], p. 14).

(4) $[\![\textbf{fake}]\!] =$
E-structure: $\lambda D_C.[\lambda x.\neg Q_E(D_C)(X) \wedge \neg Q_A(D_C)(x) \wedge \exists e_2[\textsc{Making}(e_2) \wedge \textsc{Goal}(e_2, Q_F(D_C)(x))]]$
C-structure:
$\textsc{Constitutive}$: $\lambda D_C.Q_C(D_C)$
$\textsc{Formal}$: $\lambda D_C.Q_F(D_C)$
$\textsc{Telic}$: $\lambda D_C.\neg Q_T(D_C)$
$\textsc{Agentive}$: $\lambda D_C.[\lambda x.\exists e_2[\textsc{Making}(e_2) \wedge \textsc{Goal}(e_2, Q_F(D_C)(x))]]$

With this entry for *fake*, the resulting referent will have the constitutive and formal qualia of the noun, will not have the telic or agentive qualia, and will have a new agentive qualia suggesting that the referent was made with the goal of having the same formal qualia as the noun (i.e. being a convincing fake).

In this formalism, the negation of a qualia function indicates that the function does not apply to that entity; perhaps a notation like $Q_T(D_C) = 0$ would be more natural, but I will preserve Del Pinal's notation here to avoid confusion. To compose this complex modifier with our complex noun in (2), we will need a more complex notion of function application, which Del Pinal (p. 20) provides:

(5) Dual Content Function Application (FA^{DC}):
If α is a branching node, $\{\beta, \gamma\}$ is the set of α's daughters, and $[\![\beta]\!]_E$ is a function whose domain contains $[\![\gamma]\!]$, then $[\![\alpha]\!]_E([\![\gamma]\!])$ and $[\![\alpha]\!]_C = \langle Q_C([\![\beta]\!])([\![\gamma]\!]), Q_F([\![\beta]\!])([\![\gamma]\!]), Q_T([\![\beta]\!])([\![\gamma]\!]), Q_A([\![\beta]\!])([\![\gamma]\!])\rangle$.

Per (5), the E-structure of a modifier takes in the E-structure of the noun as its argument, as does each C-structure take in its corresponding C-structure argument. Then by FA^{DC}, the result of applying [[**fake**]] in (4) to [[**gun**]] in (2) is:

(6) [[**fake gun**]] =
E-structure: $\lambda x.\neg Q_E(D_C)([\![\textbf{gun}]\!]) \wedge \neg Q_A([\![\textbf{gun}]\!])(x) \wedge \exists e_2[\textsc{Making}(e_2) \wedge \textsc{Goal}(e_2, Q_F([\![\textbf{gun}]\!])(x))]$
C-structure:
CONSTITUTIVE: $Q_C([\![\textbf{gun}]\!])$
FORMAL: $Q_F([\![\textbf{gun}]\!])$
TELIC: $\neg Q_T([\![\textbf{gun}]\!])$
AGENTIVE: $\lambda x.\exists e_2[\textsc{Making}(e_2) \wedge \textsc{Goal}(e_2, Q_F([\![\textbf{gun}]\!])(x))]$

Thus, we get a class of entities which are not guns, do not have the origins of guns, and were made to appear as if they were guns, but do not have the purpose of guns (i.e. are generally not used in shooting events). In the next section, I will provide some additional, previously unreported evidence for a Dual Content-style system, and extend the system to account for additional privative types.

3 Defending and extending the system

3.1 *Fake fake guns* and recursive privativity

Any system which does not encode a non-extension-determining part of a lexical entry, even if it invokes qualia structure, will struggle to account for recursive applications of privative adjectives. If there is only one atomic component of the lexical entry, a privative must negate it entirely. Thus, a secondary application, as in *fake fake gun*, will negate that negation, and thus return functionally the original entry for *gun*. Is this an adequate denotation for *fake fake gun*? That is, should the two *fakes* cancel each other out in this way? It is certainly true that a *fake fake gun* is probably, even necessarily, a *gun*. But it is certainly not true that every *gun* is a *fake fake gun*. The latter likely involves some element of deception, such as a criminal designing a real gun to appear like a toy in order to sneak it by security. We would not naturally call most standard-issue military or hunting weapons *fake fake guns*.

Dual Content, by contrast, produces a different meaning. Taking the denotation of *fake gun* from (6) and feeding it into the function *fake* from (4) produces the following output:

(7) $[\![\textbf{fake fake gun}]\!]$ =
E-structure: $\lambda x.\neg Q_E(D_C)([\![\textbf{fake gun}]\!]) \wedge \neg Q_A([\![\textbf{fake gun}]\!])(x) \wedge$
$\exists e_2[\textsc{Making}(e_2) \wedge \textsc{Goal}(e_2, Q_F([\![\textbf{fake gun}]\!])(x))]$
C-structure:
CONSTITUTIVE: $Q_C([\![\textbf{fake gun}]\!])$
FORMAL: $Q_F([\![\textbf{fake gun}]\!])$
TELIC: $\neg Q_T([\![\textbf{fake gun}]\!])$
AGENTIVE: $\lambda x.\exists e_2[\textsc{Making}(e_2) \wedge \textsc{Goal}(e_2, Q_F([\![\textbf{fake gun}]\!])(x))]$

Which, since the constitutive and formal qualia of *fake gun* are simply that of *gun*, can be simplified further to:

(8) $[\![\textbf{fake fake gun}]\!]$ =
E-structure: $\lambda x.Q_E(D_C)([\![\textbf{gun}]\!]) \wedge Q_A([\![\textbf{gun}]\!])(x) \wedge \exists e_2[\textsc{Making}(e_2) \wedge$
$\textsc{Goal}(e_2, Q_F([\![\textbf{fake gun}]\!])(x))]$
C-structure:
CONSTITUTIVE: $Q_C([\![\textbf{gun}]\!])$
FORMAL: $Q_F([\![\textbf{gun}]\!])$
TELIC: $Q_T([\![\textbf{gun}]\!])$
AGENTIVE: $\lambda x.\exists e_2[\textsc{Making}(e_2) \wedge \textsc{Goal}(e_2, Q_F([\![\textbf{fake gun}]\!])(x))]$

I have avoided fully simplifying the denotation in (8), since replacing all the denotation terms would simply make it more difficult to read, but reading off of this structure, we can see that a *fake fake gun* is an object which is in the extension of *gun*, has the construction and form of a gun, has the same purpose as a gun (i.e. shooting), and has the agentive quale of being made to look like something which was made to look like a gun, while in fact still being made in the way that guns are made. This last part, while a bit of a mouthful, seems to much more accurately capture our intuitions than simply equating *fake fake gun* with *gun*. This example demonstrates that DC handles complex and iterative compositions in a more natural way than non-DC systems, and elucidates an instance in which the simplicity of other systems lead them to fail to accurately capture the meaning of a term which should be straightforwardly compositional.

3.2 Pseudo-privative predicates

Classical privative adjectives like *fake* are not the only instances of privativity in our adjectival typology. Other adjectives, most commonly used in intersective or predicative ways, sometimes behave privatively. Take the case of constitutive material adjectives like *stone*. They are often simply intersective, as in *stone door*, and can be predicated of individuals, as in *That door is stone*. But they

also behave privatively, such as in *stone lion*, referring to a statue rather than a real lion. In this section, I will argue for an extension of DC to cover constitutive material adjectives, while retaining uniformity between their predicative and privative uses. Del Pinal briefly discusses constitutive material adjectives, but does not develop lexical entries nor show compositions for them. I propose that the lexical entry for *stone* is something like the following:

(9) $[\![\textbf{stone}]\!] =$
E-structure: $\lambda x.Q_C([\![\textbf{stone}]\!])(x) \wedge Q_A([\![\textbf{stone}]\!])(x)$
C-structure:
CONSTITUTIVE: $\lambda x.\textsc{Stone}(x)$
AGENTIVE: $\lambda x.\textsc{Excavated}(x) \vee \textsc{Carved}(x)$

This lexical entry is, most notably, recursive. The E-structure accesses elements of the C-structure of the same term, such that the extension of *stone* is determined the the C-structure of its members. This can be seen as a type of coercion, such that the C-structure of another term may not be extension-determining, but that adjectives that target a certain kind of qualia structure (such as *constitutive* material adjectives) may take these qualia to be necessary to satisfy their extension. We have already seen examples of promoting C-structure elements to E-structure with Del Pinal's denotation for *typical*. This denotation for *stone* is type $\langle e, t\rangle$ and so will compose with its same-typed head noun *lion* through Predicate Modification.

(10) $[\![\textbf{lion}]\!] =$
E-structure: $\lambda x.\textsc{Lion}(x)$
C-structure:
CONSTITUTIVE: $\lambda x.\textsc{Substance-Lion}(x)$
FORMAL: $\lambda x.\textsc{Perceptual-Lion}(x)$
AGENTIVE: $\lambda x.\exists e_1[\textsc{Biological-Birth-Lion}(\textsc{e}_1,\textsc{x})]$

Note that *lion* is unspecified for the telic quale. Straightforward Predicate Modification with these denotations would lead to the following denotation for *stone lion*:

(11) $[\![\textbf{stone lion}]\!] =$
E-structure: $\lambda x.\textsc{Lion}(x) \wedge Q_C([\![\textbf{stone}]\!])(x) \wedge Q_A([\![\textbf{stone}]\!])(x)$
C-structure:
CONSTITUTIVE: $\lambda x.\textsc{Substance-Lion}(x) \wedge \textsc{Stone}(x)$
FORMAL: $\lambda x.\textsc{Perceptual-Lion}(x)$
AGENTIVE: $\lambda x.\exists e_1[\textsc{Biological-Birth-Lion}(\textsc{e}_1,\textsc{x})] \wedge (\textsc{Excavated}(x) \vee \textsc{Carved}(x))$

But this is problematic. A *stone lion* should not be a lion, nor should it be composed of biological lion parts, or the result of a lion birth, and these things are explicitly contradictory with the constitutive and agentive qualia of *stone* such that the resulting set is empty, a violation of Non-Vacuity. In fact, all that

we want from *lion* is its formal quale, namely having the perceptual features or shape of a lion. To achieve this effect while preserving the predicative uses of *stone*, we will have to slightly modify our notion of Predicate Modification. To this end, I introduce the notion of *E-Precedence*. Its formal definition is given in (12), but the basic intuition is as such: whenever, in Predicate Modification, one element is specified for a part of qualia structure in its E-structure, and the other element is either unspecified for that quale or specified for it only in its C-structure, the element which has the quale in its E-structure will win out and its value for the given quale will override the C-structural value for the corresponding quale.

(12) Predicate Modification with E-Precedence (PM^{EP}):
If α is a branching node, $\{\beta, \gamma\}$ is the set of α's daughters, $[\![\beta]\!]$ and $[\![\gamma]\!]$ are both in $D_{\langle e,t\rangle}$, and $[\![\beta]\!]_E = \lambda x.Q_I(x)$ where $I \subset \{C, F, T, A\}$, then $[\![\alpha]\!]_E = \lambda x.\,[\![\beta]\!]_E\,(x) \wedge Q_J([\![\gamma]\!])(x)$, where $J \subset \{C, F, T, A\} \wedge J \cap I = \varnothing$.

In (12), I and J stand for subsets of the set of qualia. This will produce an output where the E-structure of the composed phrase includes the E-structure of β and any elements of the C-structure of γ which are not specified in the E-structure of β. Since the only element of C-structure that is specified in *lion* and not in the E-structure of *stone* is the formal, applying PM^{EP} to *stone* and *lion* gives us the following result:

(13) $$\begin{aligned}[\![\textbf{stone lion}]\!]_E &= \lambda x.\,[\![\textbf{stone}]\!]_E\,(x) \wedge Q_F([\![\textbf{lion}]\!])(x) \\ &= \lambda x.\textsc{Stone}(x) \wedge (\textsc{Excavated}(x) \vee \textsc{Carved}(x)) \\ &\quad \wedge \textsc{Perceptual-Lion}(x)\end{aligned}$$

This seems to be a good match for our intuitions: an object made of stone, produced through some kind of stonework, and with the perceptual form of a lion. This approach, then, builds in some of the basic insight of an Optimality Theory-style system for adjective composition, such as Oliver's Interpretation as Optimization (see [5]) - namely, the notion that some features are more highly-ranked than others, which will not necessarily always cause an override when they are compatible, but can do so - into the existing functional composition system. It also allows a uniform denotation for constitutive material adjectives like *stone* to capture their simple predicative and complex privative behaviors. While the introduction of this additional condition on the Predicate Modification rule is undesirable by parsimony, it does not seem possible for any solution to the problem of pseudo-privativity to be achieved without either allowing non-uniform lexical entries or some kind of modification to FA or PM, and E-Precedence seems a rather natural one following from Dual Content. This innovation will not change the analysis for true privatives like *fake*, since they compose through FA rather than PM regardless. For non-privative uses of intersective adjectives, we will simply see no clash in C-structure and therefore no override, and PM^{EP} will behave identically to standard PM.

As a final note for this section, Del Pinal observes that the word *literally* can instruct the listener to relax their commitment to Non-Vacuity and accept

seemingly empty denotations in the case of constitutive material adjectives, but not in the case of true privatives:

(14) a. Something unbelievable happened in a laboratory at Harvard. Scientists discovered a way of making, literally, stone lions.

b. Something amazing happened in a laboratory at Harvard. Some engineer managed to make, literally, a fake gun.

The observation is that we are willing to imagine a hypothetical living lion composed of stone in the case of (14a), but that we cannot do any parallel operation for (14b). In the current analysis, we could explain this by saying that *literally* is a signal for the listener to ignore E-Precedence in their interpretation. This would affect the interpretation of constitutive material adjectives, which are composed using PM, but not of true privatives, which use FA, matching the observed pattern.

3.3 Patterns of entailment

Finally, Dual Content also shows promise in covering some recent experimental data on the inference patterns of privative noun phrases. Pavlick & Callison-Burch (see [6]) show that speakers do not treat all instances of privative adjectives as entailing a negation of the head noun; specifically, speakers do still infer entailments between the privative-modified NP and the bare noun in certain cases that would not be predicted by the traditional analysis of privativity, which would predict that statements about the privative-modified NP should in fact contradict, not entail, the same statement about the bare noun. Some adjective-noun combinations 'behave in the prototypically privative way' ([6], p. 117), e.g. *counterfeit money* contradicts rather than entails *money*. Others behave contrary to this prediction, e.g. a *fake ID* is judged to still be an *ID*, a *mythical beast* is judged to still be a *beast*, and a *mock debate* is judged to still be a *debate* (or at least, statements that are true of the full NP are judged to entail the same statements about the bare noun).

The account of this data in Dual Content could follow rather naturally from the denotations for different privatives. Specifically, since different privatives negate different aspects of the C-structure, it is reasonable that you would occasionally retain entailment in some contexts if what is negated is unvalued in the noun for that particular quale. It does not seem incompatible with DC, either, that speakers would assign different relative pragmatic weights to different qualia in the C-structure, even if they aren't entirely absent.

Mock debate was judged to entail *debate* at a much greater rate than *mock execution* entailed *execution*. *Debate* probably is highly specified for its form (i.e. that it involves exchanging arguments in an oral format, most likely) but either unspecified or underspecified for its telos – perhaps to convey to the public some range of arguments, but also perhaps for competitive glory, or to convince voters to support you, but none of these seem crucial to the activity. Since *mock*, analyzed similarly to *fake*, negates the telic quale but not the formal quale,

entailment relations are likely to hold between the unmodified and modified NP in the case of *debate*. Contrast with *execution*, which is likely underspecified for its form (i.e. it may be a firing squad, gallows, guillotine, electric chair, lethal injection, or any number of less typical methods) but highly specified for its telos – it must involve ending someone's life. If *mock*, then, negates the telos but not the form, then a *mock execution* is very unlikely to hold the same entailment relations as a regular *execution*, and in fact will be considered an explicit contradiction with it.

This analysis also predicts the strangeness, possibly vacuousness, of certain privative NPs, such as *counterfeit light* (referring not to an object like a lamp, but to the light itself). *Counterfeit* is analyzed as similar but not identical to *fake*, in that it negates the agentive quale of the noun and requires that the object was made with the goal of looking and behaving like the noun ([4], p. 16). Thus, since *light* probably lacks an agentive quale, being predominately an experiential phenomenon and capable of being produced from a number of different sources and reactions, *counterfeit light* would be a strangely redundant utterance as effectively nothing is being negated. Partee's Principle of Full Interpretation (see [3]) might apply here, inducing pragmatic infelicity where *counterfeit* makes no substantive contribution to the meaning.

3.4 Conclusion: returning to the puzzle

Partee ([3]) lays out two problems for privativity: can it be accounted for in a compositional semantics, and can so-called privative nonsubsective adjectives be given a subsective denotation that explains their patterning with respect to NP splits in Polish? She argues with respect to the latter that privative adjectives coerce the noun into an expanded meaning, but the implementation of this coercion is left unspecified. I have argued that Del Pinal's Dual Content semantics offers the least theoretically costly account of privativity (adhering most closely to compositionality in the Fregean sense) while also showing the most explanatory power. By expanding the lexical entries for nouns to include an extension-determining and an associated non-extension-determining conceptual component, Dual Content allows a treatment of privative adjectives as, in a sense, subsective, since they pick out referents with some of the conceptual features of the bare noun while excluding others. This framework also allows for productive and non-vacuous analyses of adjectives like *typical*, and handles cases of iterated privativity like *fake fake guns* more successfully than its competitors, while avoiding some classical philosophical objections that other kinds of simpler lexical entries face.

I have also argued that introducing the notion of E-Precedence into Predicate Modification, such that the extension-determining components of a modifier can override the conceptual components of a noun when there is a clash, allows an extension of the Dual Content framework to account for the privative behaviors of constitutive material adjectives while preserving a uniform lexical entry for those adjectives when they are used predicatively and intersectively. This innovation integrates some of the insight of the Interpretation as Optimization theory into a

functional semantics, while avoiding its downfalls. Finally, I have suggested that Dual Content might allow for a simple explanation for some puzzling behavior of privative adjectives in recent experimental data, and that a uniform analysis of each privative adjective can still account for their inconsistent application across head nouns.

Many questions about the adjective typology still remain, including an effective way to implement non-privative non-subsective adjectives. These questions, I suspect, will be answered with tools that are perfectly compatible with, though do not depend on, Dual Content, such as degree semantics for adjectives involving time (e.g. *former*) and possible world semantics for adjectives involving possibility (e.g. *potential*). Open questions also remain about the extent of Dual Content, namely which linguistic elements need this kind of split structure and whether any other types of modification need to make reference to it. For now, this paper has shown that a functional semantics can do an excellent job of capturing our intuitions about multiple types of privativity, and even some of its less obvious behaviors, and more fully specifies what expanded denotation privative adjectives coerce their nouns into.

References

1. Kamp, H.: Two Theories about Adjectives. In E.L. Keenan, *Formal Semantics of Natural Language*, Cambribge University Press (1975)
2. Partee, B.: Privative Adjectives: Subsective Plus Coercion. In Thomas Zimmerman et al., *Presuppositions and Discourse: Essays Offered to Hans Kamp*. 273-285 (2010)
3. Partee, B.: Formal Semantics, Lexical Semantics, and Compositionality: The Puzzle of Privative Adjectives. *Philologia*. 7, 11-21 (2009)
4. Del Pinal, G.: Dual Content Semantics, privative adjectives, and dynamic compositionality. *Semantics & Pragmatics*. 8:7, 1–53 (2015)
5. Oliver, M.: Interpretation as Optimization: Constitutive material adjectives. *Lingua*. 149, 55-73 (2013)
6. Pavlick, E. & Callison-Burch, C..: So-Called Non-Subsective Adjectives. *Proceedings of the Fifth Joint Conference on Computational Semantics*. 114-119 (2016)
7. Pustejovsky, J.: The generative lexicon. *Computational Linguistics*. 17:4, 409-441 (1991)

Fighting for a share of the covers: Accounting for inaccessible readings of plural predicates

Kurt Erbach

Heinrich Heine University, Duesseldorf
erbach@uni-duesseldorf.de

Abstract. This paper presents novel empirical data that motivates an analysis of plural predicates in which the predicates have a basic, "double cover" interpretation from which all other interpretations are derived. The data presented in this paper are the results of a truth-value judgment task designed to test whether intermediate cover readings of plural predicates (i) can be made available or indexed in context as argued by Gillon [3] and Schwarzschild [9], or (ii) are never available as argued by Lasersohn [7], [8]. The results show that neither intermediate cover readings, nor collective and distributive readings are initially available in ambiguous contexts that contain minimal negative evidence. To account for the empirical data, this paper presents an analysis in which the basic reading of certain transitive constructions with two plural NPs is a Landman [6] inspired double cover reading that has been modified with a Schwarzschild [9] style approach to indexing minimal cover readings.

Keywords: Plural Predicates · Minimal Covers · Collectivity · Distributivity · Cumulativity.

1 Introduction

The interpretation of plural predicates is a still unsettled topic that draws on traditional semantic methods to motivate analyses. For example, Gillon [2] argues that plurals are ambiguous rather than vague or indeterminate in respect to readings that correspond to minimal covers of the plural noun phrase. (Gillon [2] defines a minimal cover as a set that (i) is a subset of the power-set of a set being covered, (ii) contains all of the same individuals as the set being covered, and (iii) contains no set that is a subset of another.) Lasersohn [7], however, argues that such an approach requires too many readings to be available in certain cases, and that an approach in which plural predicates are ambiguous between collective and distributive interpretations is more sound. Subsequent analyses of plural predicates fall between these two approaches, arguing for somewhere between two and (sometimes infinitely) many interpretations, e.g. [3], [4], [9], [5], [6], [8]. While there is support for each position, none of these formal analyses are informed by empirical data. In this paper, I introduce empirical data from a truth-value judgment task to motivate a new analysis of plural predicates, namely that plural predicates have a single interpretation rather than being ambiguous between two or more interpretations.

2 Background

This paper is focused on constructions like (1), in which there are two plural NPs in a transitive construction that could be interpreted as collective or distributive.

(1) Alex, Billy, and Charlie wrote songs.

(1) can be interpreted as collective, in which case Alex, Billy, and Charlie all co-wrote the same songs, and (1) can be interpreted as distributive in which case Alex, Billy, and Charlie each wrote their own songs. It is often argued that plural predicates like that in (1) are straightforwardly ambiguous between the collective and distributive readings, e.g. [7], [8], [9].

The collective and distributive readings of (1) are not the only possible interpretations, however. In addition to these interpretations, there are over 100 different combinations, or **covers** of Alex, Billy, and Charlie that could have written songs. In respect to (1), a cover is any set of sets of Alex, Billy, and Charlie, whose sum is equal to Alex, Billy, and Charlie. More formally, a cover is a subset of the the closure under sum of a set, which is equal to the supremum of the atoms of the subset.

(2) A covers B iff $A \subseteq {}^*(B) \wedge AT(\sqcup A) = B$

For example, a cover could be as complex as one in which Alex and Billy wrote songs together, while Charlie wrote songs both individually and with Alex and Billy respectively ($a \sqcup b$, c, $c \sqcup a$, $c \sqcup b$). However, while such a reading is theoretically possible, no one argues that this is part of the basic interpretation of a sentence like (1). Instead, such a sentence is argued to have a more restricted set of possible interpretations.

Gillon [2] argues that sentences like (1) are ambiguous in respect to their truth conditions, which is a set of minimal covers—i.e. sets of subsets of pluralities, in which none of the subsets overlap with the union of the others, and the union of all subsets is equal to the plurality itself (3).

(3) A minimally covers B iff A covers $B \wedge \neg\exists X(X \subseteq A \wedge \sqcup(A\text{-}X)$ covers B)

In other words, (1) has eight possible interpretations which correspond to the minimal covers of the subject NP. For example, *The men wrote musicals* is true of Rogers, Hammerstein, and Hart because, though they did not individually or collectively write musicals, the plural predicate is minimally covered by the fact that Rogers and Hammerstein wrote musicals together as did Rogers and Hart.

Lasersohn [7] criticizes the analysis of Gillon [2], claiming that certain minimal cover readings are non-existent, and that covers-based analyses are untenable because they require sentences to have unfathomably large numbers of readings. What seems to be the underlying issue for Lasersohn [7] is the distinction between *interpretation* in the sense of on-line processing of language users versus the sense of logically possible readings. For example, under Gillon's [2] analysis, (4) is predicted to be a true statement when John, Mary, and Bill are teaching assistants (TAs) who each made exactly $7,000 last year.

(4) The TAs were paid exactly $14,000 last year. [7, p. 131]

Lasersohn [7] argues that in the given context the predicted truth of this sentence is untenable. Furthermore, he argues that NPs like *the real numbers* would require infinite minimal covers and that it is unlikely that the grammar of a language would assign an infinite number of readings or set an upper limit on the number of possible readings. As an alternative to Gillon's [2] analysis, Lasersohn [7], points to analyses like Dowty [1], in which verbs are ambiguous between collective and distributive readings.

Responding to Lasersohn, Gillon [4] agrees that at least collective and distributive readings are available, but he insists that context can make available intermediate minimal cover readings—i.e. minimal cover readings other than collective and distributive. Gillon [4] gives (5-a) as an example of a context that makes intermediate cover interpretations available.

(5) a. A chemistry department has two teaching assistants for each of its courses, one for the recitation section and one for the lab section. The department has more than two teaching assistants and it has set aside $14,000 for each course with teaching assistants. The total amount of money disbursed for them, then is greater than $14,000. At the same time, since the workload for teaching a course's section can vary from one section to another, the department permits each team of assistants for a course to decide for itself how to divide the $14,000 the team is to receive.

b. The T.A.'s were paid their $14,000 last year. [4, p. 483].

While (5-a) does not explicitly point to which minimal cover is true, it nevertheless gives the context necessary to know that distributive or collective interpretations of (5-b) are not sufficient truth making conditions, and that a derivation of minimal covers is necessary.

Schwarzschild [9] also argues for a context based analysis, analyzing plural predicates as having a single meaning that can be indexed to any cover reading in the appropriate context (which solves the problem of potentially infinite covers [8]). According to Schwarzschild, [9], "whether or not a certain intermediate reading is available seems to have to do with the context not with the semantics of particular lexical items" (p. 66). He therefore proposes the following generalization to account for cover readings:

(6) $[_{S}NP_{plural}$ VP] is true in some context Q iff there is a cover C of the plurality P denoted by NP which is salient in Q and VP is true for every element in C.

This generalization for distributive readings is formalized in (7), where Part is the one place distributivity operator and Cov is free variable over sets of sets of the whole domain of quantification, the value of which is determined by the linguistic and non-linguistic context.

(7) x ∈ ‖Part(Cov)(α)‖if and only if ∀y[(y∈ ‖Cov‖∧y⊆x) →y∈ ‖α‖]
[9, p. 71]

Schwarzschild [9] specifies the translation rule in (8) which means that a plural predicate is indexed to a particular cover reading.

(8) Plural VP rule:
If α is a singular VP with translation α', then for any index i, Part(Cov_i)(α') is a translation for the corresponding plural VP.

These rules allow any cover reading to be indexed given the right context. (9-a), for example, therefore has the logical form in (9-b), where the two-place Part operation distributes the predicate to the subsets of the indexed cover, Cov_i.

(9) a. The musicians wrote songs.
b. (Part(Cov_i)(wrote'))(songs')(the-musicians')

Schwarzschild [9] concludes that the absence or presence of a given cover interpretation depends, to some extent, on the same sorts of things that other pragmatic phenomena depend on, like salience. In an ambiguous context, collective and distributive readings are made salient by the plural noun phrase.

Lasersohn [8] revisits these issues and further argues for the unavailability of intermediate cover readings, motivating an analysis where plural predicates are ambiguous between collective and distributive interpretations. While Lasersohn [7] convincingly argues that certain intermediate cover readings are never salient, it is nevertheless the case that they are logically possible interpretations.

Landman [6] takes an approach in which cover interpretations are neither one of several basic interpretations nor are they indexable via context. For Landman [6] cover interpretations are the result of a special contextual mechanism that weakens the interpretations of verbs. In respect to a plural argument like *the musicians* that denotes three individuals Alex, Billy, and Charlie or a⊔b⊔c, a minimal cover like Alex and Charlie, and Billie and Charlie (a⊔c,b⊔c in (11)), can be the agent of a plural predicate, e.g. (12)[1], so long as one has a definition of cover roles (13), a definition of covers (14), and a type shifting principle for verbs that allows verbs with plural roles to be turned into cover roles (15).

$$\begin{aligned}&\{\text{a} \sqcup \text{c}, \text{b} \sqcup \text{c}\} \in {}^*\text{MUSICIAN}\\ &[\![\text{the musicians}]\!] = \sigma({}^*\text{MUSICIAN}) = \sqcup\{\text{a} \sqcup \text{c}, \text{b} \sqcup \text{c}\} = \text{a} \sqcup \text{b} \sqcup \text{c}\end{aligned} \tag{11}$$

$$[\![\textit{The musicians wrote songs}]\!] = \begin{cases} \exists e \in {}^*\text{WRITE}: \\ a \sqcup b \sqcup c = \sigma({}^*\text{MUSICIAN}) \wedge \\ {}^C\text{Ag(e)} = \uparrow (a \sqcup b \sqcup c) \wedge \\ \exists y \in {}^*\text{SONG} \wedge {}^C\text{Th(e)} = \uparrow (y) \end{cases} \tag{12}$$

[1] AT(d) is the set of atoms below d: if d ∈ D then AT(d) = {a∈AT:a⊒d}

Let R be a thematic role

^{C}R, the cover role based on R,

is the partial function from D_e to D_d defined by: (13)

^{C}R(e) = a iff a ∈ ATOM ∧ ⊔({↓ (d) ∈ SUM: d ∈ AT(*R(e))}) = ↓(a)

undefined otherwise [6, p. 210]

group β is a subgroup of α iff $\downarrow(\beta) \sqsupseteq \downarrow(\alpha)$.

Let X be a set of subgroups in group α. (14)

X covers α iff $\sqcup\{\downarrow(x) \in X\} = \downarrow(\alpha)$ [6, p. 211]

$\lambda x_n ... \lambda x ... \lambda x_1 .\{e \in$ *V:...*R(e)=x...$\} \rightarrow$

$\lambda x_n ... \lambda x ... \lambda x_1 .\{e \in$ *V:...^{C}R(e)=x...$\}$ [6, p. 211] (15)

For Landman [6], cover readings are those in which there are plural agents of sums of events. Such readings are made possible by cover roles, which are defined in (13). If the plural role R has atoms d, and those atoms can be type-shifted down with the operation ↓, and we can take the sum of those type-shifted individuals, and that sum of type-shifted individuals is equal to the plural individual made from the group a, then a is a cover role. More plainly, if the agent of an event is a sum of groups, then that agent is a cover role. This is exactly what occurs when a sentence like (16-a) is used to describe the event that is described in (16-b)—i.e. an event in which $a \sqcup c$ and $b \sqcup c$ are the agents of separate song writing events.

(16) a. The musicians wrote songs.
b. Alex and Charlie wrote songs together, and Billie and Charlie wrote songs together.

In order to derive the interpretation in (12) from that of (16-b), the following must occur: $\uparrow(a \sqcup c)$ and $\uparrow(b \sqcup c)$ must be group atoms (made via the type shifting operation $\uparrow$[2]) that are the agents of events e and f respectively (17).

(17) $\uparrow(a \sqcup c) = \text{Ag(e)}$
$\uparrow(b \sqcup c) = \text{Ag(f)}$.

The plural agent of the sum of events e and f is equivalent to the sum of the groups $\uparrow(a \sqcup c)$ and $\uparrow(b \sqcup c)$:

(18) $*\text{Ag}(e \sqcup f) = \uparrow(a \sqcup c) \sqcup \uparrow(b \sqcup c)$ [6, p. 212]

The set of atoms below the plural agent in (18) is the set containing the two groups $\uparrow(a \sqcup c)$ and $\uparrow(b \sqcup c)$:

[2] one function of the type shifting operation ↑ is to turn plural individuals into group atoms; see [6] for details

(19) $\mathrm{AT}(*\mathrm{Ag}(e \sqcup f)) = \{\uparrow (a \sqcup c), \uparrow (b \sqcup c)\}$ [6, p. 212]

Given the definition of cover roles, (13), it is possible to take the closure under sum of the set of atoms below the plural agent, and therefore get the supremum of the groups of agents ((20)), which upshifted, is equivalent to the plural agent of events e and f ((21)).

(20) $\sqcup\{\downarrow(d)\colon d \in \mathrm{AT}(*\mathrm{Ag}(e \sqcup f))\} = \sqcup\{a \sqcup c, b \sqcup c\} = a \sqcup b \sqcup c$

(21) $*\mathrm{Ag}(e \sqcup f) = \uparrow (a \sqcup b \sqcup c)$

The type-shifting principle for verbs, (15), allows the basic meaning of the verb *write* to be shifted cover interpretations:

(22) *write* $\rightarrow \lambda y \lambda x.\{e \in *\mathrm{WRITE}\colon^{C} \mathrm{Ag}(e)=x \wedge {}^{C}\ \mathrm{Th}(e)=y\}$

This derivation provides a cover agent for the interpretation of (12) from the interpretation of (16-b).

While Landman [6] provides this mechanism for building plural predicates from covers, he argues that these are special cases that are not part of the basic interpretation of the verb. Instead, he argues there are four scopeless readings (double collective, collective-distributive, distributive-collective, and double-distributive–i.e. cumulative) if plural noun phrases fill the roles of the verb, and five other readings are available depending on how a particular scope mechanism is invoked. The cumulative interpretation is relational–i.e. it is not a statement about each individual denoted by the arguments of a transitive verb, and it is not about a predicate and one argument: it is about the relation between the predicate and its arguments. The cumulative reading (16-a) indicates that (i) there is a set of musicians, (ii) there is a set of songs, (iii) every one of the musicians wrote at least one of the songs, and (iv) every song was written by one or more of the musicians. The cumulative interpretation can be type-shifted to the "double cover interpretation", from which minimal cover interpretations can be derived, meaning that a relation between subgroups is expressed rather than a relation between individuals.

Among all of the arguments for one analysis or another, it seems that no empirical investigation into readings of plural predicates has been undertaken. Given there is no consensus among theories, it is an open question whether (i) cover readings might not be initially available but can be made available by context [4], [9], [6] or (ii) certain cover readings are never available [7],[8].

3 Main Data

In addition to distributive and collective readings of plural predicates, lexical modifiers like *each* have a distributive effect, and modifiers like *together* have a collectivizing effect [3], [9], [10]. These lexical modifiers can therefore be used to restrict the possible interpretations to distributive, (23-a), or collective, (23-b).

(23) a. Alex and Billie wrote songs individually.

b. Alex and Billie wrote songs together.

If plural predicates like *wrote songs* have all minimal cover readings available as argued by Gillon [2], then (1) should be equally ambiguous in respect to the combinations of song-writers listed in (24).

(1) Alex, Billie, and Charlie wrote songs.

(24)
a. $a \sqcup b \sqcup c$
b. $a \sqcup c, b \sqcup c$
c. $a \sqcup b, b \sqcup c$
d. $a \sqcup b, a \sqcup c$
e. $c, a \sqcup b$
f. $b, a \sqcup c$
g. $a, b \sqcup c$
h. a, b, c

If all minimal cover readings are equally available, then it should be possible to refer to a subset of the minimal covers by adding lexical modifications. For example, (25-a) is true of a set of minimal covers, and (25-b) is true of a subset of those minimal covers.

(25) a. Alex, Billie, and Charlie went to the music studio. The musicians wrote songs.
b. Alex and Billie didn't write songs individually.

The set of minimal covers that could be true of both (25-a) and (25-b) are all of those in which the predicate does not distribute to either Alex or Billie individually. The only available interpretations would be those in which Alex and Billie are part of a collective interpretation. The potentially true minimal covers are listed in (26), along with the false minimal covers, which are crossed out.

(26)
a. $a \sqcup b \sqcup c$
b. $a \sqcup c, b \sqcup c$
c. $a \sqcup b, b \sqcup c$
d. $a \sqcup b, a \sqcup c$
e. $c, a \sqcup b$
f. ~~$b, a \sqcup c$~~
g. ~~$a, b \sqcup c$~~
h. ~~a, b, c~~

It is also possible to use modifiers to eliminate collective interpretations for particular individuals. In (27) for example, the use of *together* in (27-b) negates the scenarios in which Alex and Billie are predicated over collectively.

(27) a. Alex, Billie, and Charlie went to the music studio. The musicians wrote songs.
b. Alex and Billie didn't write songs together.

The set true and false minimal covers for (27-a) and (27-b) are listed in (28)[3].

[3] though $p \sqcup q$ is only a subpart of $p \sqcup q \sqcup r$, this reading is assumed to be canceled via implicature

(28)
a. ~~$a \sqcup b \sqcup c$~~
b. $a \sqcup c, b \sqcup c$
c. ~~$a \sqcup b, b \sqcup c$~~
d. ~~$a \sqcup b, a \sqcup c$~~
e. ~~$c, a \sqcup b$~~
f. $b, a \sqcup c$
g. $a, b \sqcup c$
h. a, b, c

Taking these modifications one step further, only a single minimal cover is available as the truth-making condition when using both *each* and *together* in the same sentence. For example, given (29-a) as a context, (29-b) negates all minimal covers in which *wrote songs* gets a collective or distributive interpretation in respect to Alex and Billie.

(29)
a. Alex, Billie, and Charlie went to the music studio. The musicians wrote songs.
b. Alex and Billie didn't write songs individually or together.

Both (29-a) and (29-b) are true if Alex and Charlie wrote songs together and Billie and Charlie also wrote songs together. The true and false minimal covers of these two sentences are listed in (30).

(30)
a. ~~$a \sqcup b \sqcup c$~~
b. $a \sqcup c, b \sqcup c$
c. ~~$a \sqcup b, b \sqcup c$~~
d. ~~$a \sqcup b, a \sqcup c$~~
e. ~~$c, a \sqcup b$~~
f. ~~$b, a \sqcup c$~~
g. ~~$a, b \sqcup c$~~
h. ~~a, b, c~~

Given the interpretation of plural predicates is an open question, there are five ways in which the pairs of sentences in (25), (27), and (29) are likely to be interpreted. If all of these follow-up sentences are judged to be possibly true, then it could be the case that the plural predicates are straightforwardly ambiguous between all minimal cover interpretations as argued in Gillon's [2], [3] earlier work, or it could be the case that plural predicates are ambiguous between collective and distributive interpretations, and that context makes the minimal covers available as argued by Gillon [4] and Schwarzschild [9], and implied by Landman [6]. Second, if (25) and (27) are judged to be possibly true, and (29) is judged to be necessarily false, then it could be the case that plural predicates are ambiguous between distributive and collective interpretations but intermediate cover interpretations are not available, as argued by Lasersohn [8]. Third, if (25) is judged to be possibly true while (27) and (29) are judged to be necessarily false, then it would be the case that a collective interpretation is basic and all other interpretations are derived. Fourth, if (27) is judged to be possibly true while (25) and (29) are necessarily false, then the distributive interpretation is basic and all other interpretations are derived. Lastly if all follow-up sentences are judged to be false, then it is the case that there is a single general interpretation that is basic, and all other interpretations are derived or indexed.

Experimental Design. An empirical study was designed to test determine the interpretations of the pairs of sentences, like those in (25), (27), and (29). A truth-value-judgment survey was conducted with 32 native English speakers through Prolific.ac. The participants were presented with 45 test items containing

a context like (29-a) and a follow-up like (29-b). Participants were told to judge whether the follow-up sentence could be true or must be false in respect to the context preceding it[4]. The 45 test items exemplified one of the three conditions in (25), (27), and (29): 15 test follow-up items contained *individually*, 15 contained *together*, and 15 contained both *individually* and *together*. Participants were also asked to judged the truth value of 45 filler items that could be true or must be false depending on their lexical modifiers. The total number of items expected to be true or false was equal.

Results. The results of the study show that there is a significant difference in the way that the truth of sentences with both *individually* and *together* are judged relative to sentences with only one of the two lexical modifiers. Using a binary logistic regression model (lme4 package in R), and the conditions and judgments as arguments, the judgments of test condition with both *individually* and *together* were found to be significantly different ($p<0.001$) than judgments of the condition in which sentences only contained *together* as a lexical modifier. Sentences that only contained *individually* as a lexical modifier were found to be judged no differently ($p=0.282$) than those that only contained *together*. These results show that despite the fact that each follow up sentence is true in respect to its preceding context, speakers do not judge sentences in the test condition to be true at the same rate at which they judge sentences in the other conditions to be true.

The average percentage of true and false judgments for sentences in each condition is presented in Figure 1. This graph shows that follow up sentences with

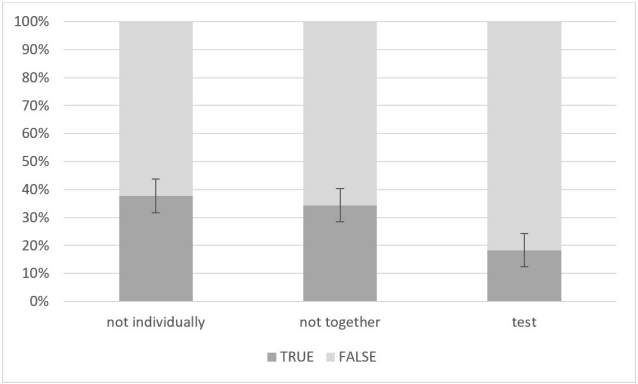


Fig. 1. Average percentage of true and false judgments by condition

only one of the two lexical modifiers are judged as necessarily false a majority of the time, while follow up sentences with both lexical modifiers are judged

[4] While these directions were written above every pair of sentences, the options the participants clicked on were simply labeled *True* and *False*.

as false an even larger majority of the time. In other words, negated follow up sentences that restrict the set of true minimal covers with the lexical modifiers *individually* or *together* are generally judged to be false. This is a surprising result given the plural predicates are said to have both collective and distributive readings, yet neither reading seems to be available when the subjects were asked to interpret the possible truth of follow-up sentences. If the collective reading was available, then the follow-up sentences negating the distributive reading should all have been true. Furthermore, if the distributive reading was available, then the follow-up sentences negating the collective reading should have been true.

Discussion. The fact that the follow-up sentences were judged to be false means that the plural predicate they follow is not straightforwardly ambiguous between all minimal covers as argued for by Gillon's earlier work [2],[3]. It also cannot be the case that they are ambiguous between collective and distributive interpretations argued by Lasersohn [8], Schwarzschild [9] and Gillon's later work [4]. The results also suggest that the follow-up sentences in the study are insufficient context to make the set of true cover readings available. Instead of any of the aforementioned analyses, the empirical data seems to point toward an analysis in which neither the distributive, collective, nor intermediate cover readings are part of the basic meaning.

4 Analysis

Building on the idea of Schwarzschild [9] that a plural predicate has one meaning that can index cover interpretations, and also the idea from Landman [6] that cover readings are derived from a double cover interpretation, I motivate an analysis in which plural predicates have a single, general interpretation from which all cover interpretations are indexed. The double cover reading from Landman [6] provides a weak, general meaning for the plural predicate, and by adding indexing, specific interpretations can be salient. The required translation entails the following rule.

(31) If α is a singular transitive verb phrase with translation A, then for any index i, $\exists e \in {}^*\mathrm{A} : {}^{C_i}\mathrm{Ag(e)} = x \wedge {}^{C_i}\mathrm{Th(e)} = (y)$ is the translation for the corresponding plural transitive verb phrase.

If a particular cover is not indexed in the context–i.e. the index is left unspecified as i–then the plural predicate is straightforwardly interpreted as a dual cover reading. The reading indicates (i) that there is a sum of writing events, (ii) a sum of groups of musicians (Alex, Billie, and Charlie in (25), (27), and (29)) as a plural agent, (iii) there is a sum of groups of songs as a plural theme:

$$[\![\textit{The musicians wrote songs}]\!] = \begin{cases} \exists e \in {}^*\mathrm{WRITE} : \\ a \sqcup b \sqcup c = \sigma({}^*\mathrm{MUSICIAN}) \wedge \\ {}^{C_i}\mathrm{Ag(e)} = \ \uparrow (a \sqcup b \sqcup c) \wedge \\ \exists y \in {}^*\mathrm{SONG} \wedge {}^{C_i}\mathrm{Th(e)} = \ \uparrow (y) \end{cases} \tag{33}$$

While this seems very similar to a distributive interpretation (and in Landman's [6] framework, the double cover interpretation is a type-shifted double-distributive (cumulative) interpretation), without indexing a particular cover, it is impossible to tell exactly which (covers of) musicians wrote exactly which (covers of) songs. It is therefore distinct from Landman's [6] scoped distributive readings where the set of musicians would necessarily distribute to either distinct sets of songs, or the same set of songs.

The proposed analysis, provides a plausible explanation for why each condition was judged to be necessarily false in the empirical study. The ambiguous context in which the plural predicate was presented did not index any minimal cover despite the fact that it informed the participants that every atomic part of the song writing event had a group of musicians as the agent and a group of songs as the theme. The ambiguous context did not index even the strictly collective or distributive interpretation of the agent or the theme, so no specific interpretation from the set of minimal covers was available. At the same time, the follow-up sentences were interpreted as negated collective, distributive, and both collective and distributive readings respectively, these readings being indexed by the use of the lexical modifiers *together* and *individually*. Crucially, these indexed readings in the follow-up sentences were for an agent that was subset of the agent in the context sentences. Because no specific cover was indexed in the context sentences, the intersection of the context sentence and the follow-up sentence was the empty set. It seems that the participants in this study judged the follow-up sentences to be necessarily false because the follow-up sentences did not contain information that could straightforwardly index a particular cover reading of the the preceding context. In other words, the sort of context that can index a particular cover interpretation is positive evidence. The negative evidence in this study's follow-up sentences is not sufficient for indexing cover interpretations of the preceding contexts: Given Alex and Billy are part of the double cover interpretation of the context, the follow up sentences were generally judged to be false.

The fact that follow-up sentences with both *individually* and *together* were judged false significantly more frequently than those with only *individually* or *together*, is a phenomenon that must be accounted for. It might suggest that collective and distributive readings are more simple to derive than intermediate cover readings, which corresponds to the claim supported by many that these are basic readings—e.g. [3], [6], [7], [9]. However, given these readings cannot be taken to be basic readings in light of the evidence found in this study, the following question remains open: Why are collective and distributive readings more simple to get than intermediate cover readings?

One possible explanation for the difference in judgments is the respective frequencies of overtly collective, distributive, and intermediate cover readings. Both the number of lexical modifiers that specify collective or distributive readings and their frequency of use lend to the intuition that these two minimal cover readings are more salient than intermediate covers. After all, it seems there are no lexical modifiers that index specific intermediate covers, and situations in which intermediate covers are salient are likely to be less frequent than situations in

which collective or distributive interpretations are salient. A corpus study looking for the relative frequencies of these readings could validate this hypothesis.

5 Conclusion

While it is possible for plural predicates to have collective or distributive interpretations, their basic interpretation is more general. The results of the empirical study in this paper suggest that neither Gillon [3], Landman [6], Lasersohn [8], nor Schwarzschild [9] is correct in concluding that the collective and distributive interpretations are part of the basic interpretation of plural predicates. At the same time, the study also suggests that Lasersohn [8] is correct in arguing that certain intermediate cover readings are never available, that is if they are never made contextually salient. I propose a basic reading, inspired by Landman's [6] double cover reading and Schwarzschild's [9] indexing, that can index cover readings when they are contextually salient. Given this contradicts the common view, further empirical research is necessary to substantiate these claims.

Acknowledgments

Special thanks to Leda Berio, Peter Sutton, Hana Filip, and participants in the Semantics and Pragmatics Exchange and the Graduate Research Seminar at Heinrich Heine University. This research was funded by the German Research Foundation (DFG) CRC 991, Project C09.

References

1. David Dowty. Collective predicates, distributive predicates and *all*. In Proceedings of the 3rd ESCOL, pages 97–115. (Eastern States Conference on Linguistics), Ohio State University Ohio, 1987.
2. Brendan S Gillon. The readings of plural noun phrases in english. Linguistics and philosophy, 10(2):199–219, 1987.
3. Brendan S Gillon. Bare plurals as plural indefinite noun phrases. In Knowledge representation and defeasible reasoning, pages 119–166. Springer, 1990.
4. Brendan S Gillon. Plural noun phrases and their readings: A reply to lasersohn. Linguistics and Philosophy, 13(4):477–485, 1990.
5. Fred Landman. Groups, i. linguistics and philosophy, 12(5):559–605, 1989.
6. Fred Landman. Events and plurality: The jerusalem lectures. number 76 in studies in linguistics and philosophy, 2000.
7. Peter Lasersohn. On the readings of plural noun phrases. Linguistic inquiry, 20(1):130–134, 1989.
8. Peter Lasersohn. Plurality, conjunction and events, volume 55. Springer Science & Business Media, 2013.
9. Roger Schwarzschild. Pluralities, volume 61. Springer Science & Business Media, 1996.
10. Kristen Syrett and Julien Musolino. Collectivity, distributivity, and the interpretation of plural numerical expressions in child and adult language. Language acquisition, 20(4):259–291, 2013.

Representing Scalar Implicatures in Distributional Semantics

Maxime Codere Corbeil

University of Quebec in Montreal, Montreal, Canada

Abstract. I use compositional distributional models as a lens through which to examine scalar implicatures. I will look at two opposing views regarding the derivation of scalar implicatures, i.e. the localist and the globalist views, and will illustrate how they would respectively be integrated within these distributional models.

Keywords: scalar implicature, compositional distributional semantics, sentence similarity

1 Introduction

Distributional approaches to modelling meaning in use take some contextual clues into account while simultaneously lending themselves to computational implementations [2]. In such approaches, the meaning of a word is represented by a vector consisting of the relative occurrence of this word with respect to other words within a certain distance of it. Distributional models were first developed to deal with words in isolation but they rapidly expanded so they could represent not only words but whole sentences [2]. Compositional Distributional Semantics (CDS) is interested in finding the best way to compose a sentence-vector from the combination of word-vectors so that the resulting sentence-vector will in fact represent the meaning of the sentence. In standard CDS models, a sentence-vector represents the meaning of the composition of the words forming the sentence but it does not take into account the information that could be derived from this composition. The idea here is to discuss how implicated meaning would be integrated within the sentence-vector and also what would be the consequences for sentence similarity measures since the implicated meaning have an important role to play for such applications as paraphrase detection and short answer tasks [10].

One goal of this article is to bridge the gap between the distributional representation of a sentence and its meaning that may be inferred from the context. Scalar implicatures (SI) are one such type of inference that will be important for advances in this area. After a brief review of the two most prominent theories of scalar implicature, namely the localist view [1] and the globalist view [6], we will examine two compositional distributional models, Simple Multiplicative Compositional Model [13] and Category Compositional Distributional Semantics [3], and use them as a tool to represent sentence-vectors involving scalar

implicatures. We then compare the two views of SIs by looking at sentence similarity computation involving different cases of implicated meaning. Our results are then discussed in terms of what we would expect to achieve by integrating the implicated meaning sentence-vector in CDS.

2 Scalar Implicatures

Scalar implicature (SI) is a special kind of quantity implicature involving a series of alternatives that can be generated by substituting lexical items that are linked via a conceptual scale [6]. Horn defined conventionalized scales as containing different lexical items organised by informativity [8]. For example, from (1) we can derive (3) using the scale in (2):

(1) John ate some candies.
(2) {*some, all*}
(3) John ate some and not all candies.

There are many different theoretical models of SI and its derivation. In this paper, we will focus solely on the opposition between what we call globalist models and localist models, looking only at the differences most relevant to their implementation in distributional models.

2.1 Globalist view

The globalist view states that SIs are always derived using pragmatic mechanisms [6]. To understand the standard derivation procedure, suppose that Mike utters (4).

(4) John ate some candies.
(5) John ate all the candies.
(6) John ate some and not all candies.

(i) Mike utters (4) instead of (5)
(ii) He must not believe that (5) is true: $\neg B_M(5)$ (read: M does not believe (5))
(iii) Mike is likely to know whether (5) is true: $B_M(5) \vee B_M\neg(5)$.
(iv) Combining (ii) and (iii) we get that Mike must believe (5) is not true: $\big(\neg B_M(5) \wedge (B_M(5) \vee B_M\neg(5))\big) \rightarrow B_M\neg(5)$ which gives rise to (6).

As shown, the hearer first derived the weaker implicature ($\neg B_M(5)$) based on the fact that the speaker chose not to utter the stronger alternative in (5). The transition from the weaker implicature to the stronger, i.e. $B_M\neg(5)$, is made possible through the Competence Assumption wherein a competent speaker either believes p or $\neg p$ [6]. It is thus the combination of the Competence Assumption and the choice made by Mike to utter the weaker alternative that leads to the derivation of a SI.

This model is considered globalist because SIs are derived using global processes like contextual information and the Competence Assumption rather than

local grammatical processes. This view also supports the idea that SIs are never derived by default; rather, they are derived in particular contexts by post-compositional global processes, which means that the derivation of SIs should not play a role in the compositional process.[1]

2.2 Localist view

In contrast to the Globalist view, the localist view considers that SIs are derived from specific lexical items. Additionally, and most importantly, SIs are derived compositionally, which means they are derived semantically and not pragmatically [1]. The basic assumption here is that the implicature arises from the presence of an exhaustivity operator O that may be expressed in some cases as a silent covert *only*.[2] Going back to our example *John ate some candies*, the steps to go from (4) to (6) in the localist view are as follows:

(i) The set of alternatives of (4) is based on the Horn scale ALT((4))={*some*, *all*}
(ii) The logical form of (4) is then added to the negation of all stronger alternatives of (4): $O_{ALT}(4) = some \wedge \neg all$
(iii) The scalar implicature is thus derived gramatically from the presence of the operator O, i.e. *John ate O some candies* that negates the stronger alternatives and this gives (6).

The fact that this operator acts locally implies that it is possible to retrieve different implicatures depending on the site where the operator is processed [1]. After introducing the details of compositional distributional models in the next section, we will see in section 4 that these different derivations of SIs will also lead to different treatments by compositional models, and this, even for non-embedded cases. In this paper, we will only consider cases where the scalar implicatures are in fact derived and will not discuss cases where they are not.[3]

3 Compositional Distributional Semantics

The main idea of Compositional Distributional Semantics (CDS) is to start from the word-vectors and combine them together to obtain a sentence-vector representing the meaning of the complete sentence. In this paper,we present two CDS models: Simple Multiplicative Compositional Model (SMCM) [13] and Category Compositional Distributional Semantics (CCDS) [3]. The former uses simple operations between vectors in its composition and the latter is able to derive sentence meaning while taking into account the structure of the composed sentence.

[1] Geurts [6] acknowledges the fact that some kind of local pragmatics must be invoked when considering particular examples such as embedded implicatures, but those structures are outside the scope of this article.
[2] This view originates from the treatment of embedded implicatures, but also extends to all scalar implicatures [5]. For more information about the motivations behind the localist view see [1,5].
[3] For a discussion about the optionality of scalar implicature see [6,5].

3.1 Mitchell and Lapata

We present here only the multiplicative model of Mitchell and Lapata since they arrived at the conclusion that "simple multiplication" is better suited for sentence similarity tasks [13, p.1417] than all the other alternatives they tested.

(7) Multiplicative Model: $p_i = u_i \cdot v_i \cdot w_i$

The multiplicative function defines the ith component of the sentence-vector as being composed from the simple multiplication of the ith components of the word-vectors $\overrightarrow{u}$, $\overrightarrow{v}$ and $\overrightarrow{w}$. This function is commutative, which means that the order the word are combined in does not change the result, meaning that syntactic structure is not taken into account.

3.2 Coecke et al.

The idea behind this approach is that "syntax drives the compositional process" [2, p.26]. Their compositional approach has 3 steps [3]. The first step is about assessing the syntactic constraints, the second one is about linking grammatical analysis with semantics, and the third step is about the composition itself.

(i) Assign a grammatical type p_i to each word w_i of a string of words then apply the axioms and rules of pregroup grammar to reduce these types [11]. Following other categorial grammars, a sentence has the type s, nouns like *John* and *candies* are assigned grammatical type n, determiners like *some* are assigned nn^l because they combine with a noun, and transitive verbs such as *ate* are assigned the grammatical type $n^r sn^l$ because they take n as input for subject and object and output a type s of a sentence.[4] To reduce the form for the sentence we used the cancellation rules of pregroup grammar which state that any type X will cancel if combined to the left of its right adjoint X^r, and any type X will cancel if combined to the right of its left adjoint X^l, i.e. $XX^r \to 1$ and $X^lX \to 1$ where 1 is the Identity. Put together, it would look like this:

(8)
$$\begin{array}{cccccc} \text{John} & \text{ate} & \text{some} & \text{candies} \\ n & \underbrace{n^r s n^l} & \underbrace{n n^l} & n \\ & \underbrace{1 s n^l} & & \underbrace{n 1} \\ & & \underbrace{\qquad\qquad s} \end{array}$$

(ii) Assign a vector space for every syntactic type present in the sentence.

(iii) Combining the vectors of the meaning for every word in the spaces built above, we get the representation of the sentence-vector where $^\intercal$ is the transpose and $\times$ is the matrix multiplication:[5]

$$\overrightarrow{\text{John ate some candies}} = \overrightarrow{John}^\intercal \times ate \times (\overrightarrow{some}^\intercal \times \overrightarrow{candies}) \qquad \text{(Eq. 3.1)}$$

[4] For simplicity reasons, all nouns are treated as grammatical type n [7]. See Lambek [11] for a richer description of basic types.

[5] The details of the calculations are presented in [3] and [7].

4 Representing SI Using CDS

Now that we have introduced two CDS models we can go back to SIs and discuss how these models could be used to represent the sentence-vector resulting from the SI. For the localist view, the SI sentence vector would look like this, where F represents a compositional function combining all the words together:

$$\begin{aligned}\overrightarrow{SI_{LOC}} &= F(John, ate, O, some, candies) \\ &= F(John, ate, some, \text{and not all}, candies) \\ &= \overrightarrow{\text{John ate some and not all candies}}\end{aligned} \qquad \text{(Eq. 4.2)}$$

The fact that the exhaustification operator *Only* acts locally at the compositional stage makes it straightforward to compute the SI sentence-vector for the localist view since it only requires substituting *some* with *some and not all* in the original sentence. In the localist case, the composition happens only once because the negation of the stronger alternative is integrated within the sentence before the sentence is composed. In turn, the SI sentence-vector is thus equivalent to the vector for the composed sentence *John ate some and not all candies*. To derive the representation for the meaning of the SI sentence-vector we just have to choose which compositional function to use depending on the CDS model.

The globalist view, on the other hand, considers that the implicated information *and not all* is processed post-compositionally. This means that the original sentence *John ate some candies* must be already composed before the new implicated information is integrated within the meaning of the sentence. Following the Gricean framework, the original sentence is equivalent to the meaning of "What is said" while the implicated information corresponds to the meaning of "What is implicated".

(9) What is said — *John ate some candies* | What is implicated — *John ate not all candies*

One key difference between the localist and the globalist views is that in the globalist view the fact that "What is said" is already composed implies that the implicated meaning could not possibly only be *and not all* because it is not propositional by itself. Instead the implicated meaning should be a complete proposition containing the same information, namely *John ate not all candies*. In the globalist view, the complete interpretation of the sentence *John ate some candies* must integrate both the "What is said" and the "What is implicated" components. We thus have to first compose the original sentence ("What is said") and the SI sentence ("What is implicated") separately. Once these two sentences are formed, we may then compose them together.

$$\begin{aligned}
&\overrightarrow{\text{Complete meaning}_{GLO}} = F(\overrightarrow{\text{What is said}}, \overrightarrow{\text{What is implicated}}) \\
&= F(\overrightarrow{\text{John ate some candies}}, \overrightarrow{\text{John ate not all candies}}) \\
&= F_3\Big(F_1(John, ate, some, candies), F_2(John, ate, not, all, candies)\Big)
\end{aligned} \tag{Eq. 4.3}$$

In (Eq. 4.3) the indices of the compositional function F are only there to explicitly show that there is three different compositions. This complete sentence-vector derived using the globalist view is equivalent to the SI sentence-vector we derived from the localist view. Equivalent in the sense that they both represent the same conveyed meaning about the fact that John ate some candies but that he did not ate all of them. With this in mind we can go back to the two CDS models we presented and compute the resulting vectors.

4.1 SI and SMCM

Substituting the multiplicative function "$\cdot$" for F we can derive the final vector for the localist view and the globalist view using SMCM:

$$\begin{aligned}
&\overrightarrow{\text{Complete meaning}_{LOC}} \\
&\qquad = \overrightarrow{\text{John}} \cdot \overrightarrow{\text{ate}} \cdot \overrightarrow{\text{O some}} \cdot \overrightarrow{\text{candies}} \\
&\qquad = \overrightarrow{\text{John}} \cdot \overrightarrow{\text{ate}} \cdot \overrightarrow{\text{some}} \cdot \overrightarrow{\text{and}} \cdot \overrightarrow{\text{not}} \cdot \overrightarrow{\text{all}} \cdot \overrightarrow{\text{candies}}
\end{aligned} \tag{Eq. 4.4}$$

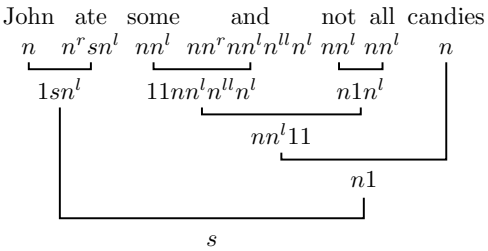
$$\begin{aligned}
&\overrightarrow{\text{Complete meaning}_{GLO}} \\
&= F_3\Big(F_1(John, ate, some, candies), F_2(John, ate, not, all, candies)\Big) \\
&= \Big(\overrightarrow{John} \cdot \overrightarrow{ate} \cdot \overrightarrow{some} \cdot \overrightarrow{candies}\Big) \cdot \Big(\overrightarrow{John} \cdot \overrightarrow{ate} \cdot \overrightarrow{not} \cdot \overrightarrow{all} \cdot \overrightarrow{candies}\Big)
\end{aligned} \tag{Eq. 4.5}$$

This suggests that even in commutative CDS models that do not involve syntactic constraints, the resulting vector for the localist view and for the globalist view are likely to be different. This result might be surprising but it stems from the fact that globalist view is about composing two sentences together which amounts for duplicates of certain words like *John*, *ate* or *candies*. It would make no sense to treat the SI as only being the particle *and not all* because it is neither a sentence nor propositional. The fact that SIs are derived post-compositionality thus trigger this difference between the localist and globalist views when using this CDS model. Here $\overrightarrow{\text{Complete meaning}_{LOC}}$ and $\overrightarrow{\text{Complete meaning}_{GLO}}$ respectively represents the sentence-vector for the SI for the localist view and the combination of "What is said" and "What is implicated" for the globalist view.

4.2 SI and CCDS

When considering the CCDS approach to composition, grammatical types and word order matter. The compositional process of pregroup grammar works fine if we simply substitute *some* with *some and not all* as long as the correct grammatical types are chosen for every word. The conjunction *and* has a type $x^r x x^l$ and here $x = nn^l$.

(10)



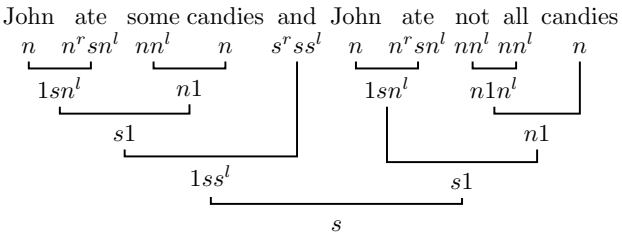
From this we can compute the resulting vector, where $\times$ is a matrix multiplication and $\odot$ is a point-wise product also known as Hadamard product:

$$\overrightarrow{\text{John ate some and not all candies}} = \overrightarrow{John}^{\intercal} \times \overrightarrow{ate} \times \left[\left(\overrightarrow{some} \odot \left(\overrightarrow{not} \times \overrightarrow{all} \right) \right) \times \overrightarrow{candies} \right] \tag{Eq. 4.6}$$

To arrive at this result we used the simplification presented in [9]. Since a conjunction always outputs the same type as its conjuncts and since it "enforces equal contributions of the conjuncts in the final result" [9, p.33] it can be replaced by point-wise product of its conjuncts.

Coming back to the globalist view, the basic idea remains the same but now we must first compose "What is said" and "What is implicated" separately before being able to compose them together. Once we have computed both components of the complete meaning, the idea is to compose them together using a coordination relation. The conjunction *and* is the simplest way to combine two sentences together and here *and* has the grammatical type $(s^r s s^l)$, i.e. s instantiates x from $x^r x x^l$. To make sure the complete sentence will compose we can verify the type of the whole sentence.

(11)



As described before, we can simplify the conjunction *and* by using a pointwise product assuring that the two conjuncts contribute equally to the output of the sentence.

$$\overrightarrow{\text{Complete meaning}_{GLO}} = \left(\overrightarrow{John^{\dagger}} \times \overrightarrow{ate} \times (\overrightarrow{some} \times \overrightarrow{candies})\right) \odot \left(\overrightarrow{John^{\dagger}} \times \overrightarrow{ate} \times \left((\overrightarrow{not} \times \overrightarrow{all}) \times \overrightarrow{candies}\right)\right) \quad \text{(Eq. 4.7)}$$

We now have computed the sentence-vectors for both the localist and the globalist views using two distinct CDS models and we arrived both times at different results. In the next section we will quantify this difference using sentence similarity measures.

5 SI and Sentence Similarity

Vectorial approaches to meaning have the advantage of facilitating the comparison between meaning because it is mathematically easy to compare two vectors. The most used measure for the distance between two vectors is the cosine similarity measure. The cosine measure is the computation of the angle between the two vectors. In (Eq. 5.8) $\|\overrightarrow{w}_1\| = \sqrt{\sum_i w_i^2}$ is the norm of the vector. The more similar two vectors are, the smaller the angle will be, where an angle of 0 would yield a perfect similarity of 1.

$$Sim(\overrightarrow{w}_1, \overrightarrow{w}_2) = Cosine(\overrightarrow{w}_1, \overrightarrow{w}_2) = \frac{\overrightarrow{w}_1.\overrightarrow{w}_2}{\|\overrightarrow{w}_1\|\|\overrightarrow{w}_2\|} \quad \text{(Eq. 5.8)}$$

The similarity measures when using SMCM and CCDS are presented in Table 1. For our calculations, we used pre-trained vectors from the Wikinews corpus [12].[6] We first computed the similarity between the original sentence *John ate some candies* and the two complete sentence-vector we derived from the localist and the globalist view. To make sure our results did not overtly depends on the meaning of the subject *John* and the object *Candies*, we varied the subject and the object and took the average value.[7]

		Cosine	s_{cos}
SMCM	Localist - Base sentence	0.9549	0.0016
	Globalist - Base sentence	0.8863	0.0070
CCDS	Localist - Original sentence	0.9927	0.0001
	Globalist - Original sentence	0.8709	0.0025

Table 1: Cosine similarity measures between complete and base sentence-vector

[6] These vectors were trained using *fastText*, have 300 dimensions and are available for download at https://fasttext.cc/docs/en/english-vectors.html

[7] 3 subjects: *John*, *Mary*, *Someone*; 3 objects: *candies*, *cookies*, *apples*; 9 possible permutations.

The first thing we note from the results presented in table 1 is that the standard deviation is very small which tends to show that the similarity measures we obtained depend more on the structure of the derivation process for the SI that on the individual values of the words that composed them. We can also see that the localist sentence-vector is closer than the globalist sentence-vector to the original sentence-vector by a significant margin. Under the CCDS model, the localist sentence-vector is even closer to the original one. In fact it is so close we could even say it is almost equivalent to the vector representation of the original sentence *John ate some candies*.

5.1 Other Kinds of Quantity Implicatures

In this paper we have focussed on SIs of the form *some and not all*, but there are a multitude of different kinds of quantity implicatures [6]. To see if our results would also hold for other kinds of SI we also computed similarity measures for cases when *John ate 3 candies* lead to the derivation of the following implicated meaning: *John ate no more than 3 candies*. We also computed the similarity values for a kind of quantity implicatures called free choice inference [5,6]: *John or Mary ate candies* leading to the derivation *John or Mary and not both ate candies*. Similarly to *no more than*, the results are much lower for the globalist view than for the localist view.

			Cosine
no more than	SMCM	Localist - Original sentence	0.8571
		Globalist - Original sentence	0.7398
	CCDS	Localist - Original sentence	0.9178
		Globalist - Original sentence	0.7421
and not both	SMCM	Localist - Original sentence	0.9491
		Globalist - Original sentence	0.6554
	CCDS	Localist - Original sentence	0.9861
		Globalist - Original sentence	0.8017

Table 2: Similarity for *no more than* SIs and *and not both* free choice inferences

Our results are still in line with what we got for *some and not all*. In the case of *no more than*, the relative scores for similarity measures remain the same as before, i.e. the localist view is still the one closer to the original sentence, but this time the globalist view sentence-vector is even more dissimilar than before. We note there are no clear discrepancies in our results between the two compositional models as respective trends of values is constant throughout the results.

6 Discussion

The results presented here clearly point toward the localist view to be more similar to the original sentence. The question now is whether these results say something about the difference between the localist and the globalist view.

6.1 Length of the sentence-vector

Even though it would be difficult to measure precisely the contribution from the meaning of the words we can at least convinced ourselves that it is not completely overruled by the difference in word number by comparing the similarity measures between different sentences having the same length. We thus computed similarity measures by respectively varying 3 words and 6 words while keeping the same first 4 words throughout as in our other examples for SI. Using SMCM and 4 different versions for the added words, we got that the ratio of similarity measures between the 7-word sentence and the 10-word sentence varied by approximately 7.5%. These results tend to show that the meaning of the individual words are in fact making a difference for sentence similarity and that they are not just drowned out by the difference of length of the sentence-vectors.[8]

6.2 Contribution of the Implicated Meaning

The 'SI meaning' is responsible for the difference in similarity between the complete sentence and the original one. A lower similarity score thus means that the SI part contributes more to the complete meaning sentence-vector whereas a higher similarity score would imply the contrary.

We must first discuss the fact that the contribution of the implicated meaning to the complete sentence-vector is generally so low for the localist view that it could be a problem because we could question whether the SI-part was taken into account at all while this issue would not be raised when using the globalist view. In theory, if the goal of sentence similarity measures is to compute the contribution from the SI part, then the localist view might not be the best candidate because its SI contribution seems drowned out by the composition of the sentence and this results in the minimization of its contribution.

6.3 Short Answer Tasks

Short answer scoring tasks are about determining if the answer provided by the student is valid or not compared to the given target answer. According to [10] paraphrase detection consisting of extracting similar sentences from corpora is a method of choice for short answer scoring tasks. If the goal of sentence similarity models is to measure the similarity between pairs of sentences as accurately as possible, then it should be just natural to want to integrate this complete similarity within sentence similarity measures.

(12) *Question:* Did John steal all the candies?

(13) *Target A:* No John did not steal all the candies, but he did steal some.

(14) *Target B:* No John did not steal all the candies.

(15) *Student #1 answer :* John stole some candies.

[8] It is important to note this analysis is merely qualitative, but at this point we are mostly interested by a qualitative analysis of the difference between the localist and the globalist view.

The Cosine value between (13) and the complete sentence-vector of (15) is 0.8148 for the localist view and it is 0.8845 for the globalist view under SMCM; while the Cosine value between (14) and the complete sentence-vector of (15) is 0.9887 for the localist view and it is 0.9563 for the globalist view under SMCM.

The similarity values varies according with what we would expected but it seems the localist view seems to be more selective overall. When using Target B the localist view is only 3.3% higher than the globalist one, but it is 7.9% lower than the globalist view when using Target A. To illustrate this let us chose the threshold of acceptance, i.e. the similarity value below which the student's answer would be rejected, to be 0.85. Then the globalist view would accept the student's answer in both cases, while the localist view would only accept the answer when compared with Target B because its form is much closer to the localist complete sentence-vector.

6.4 Translation Studies and Sentence Similarity Measures

Although the integration of the implicated meaning within the complete sentence-vector is not really discussed in DS right now, some other fields already emphasise the important difference between original sentence meaning and implicated meaning. To illustrate this difference we borrow an example described in [4] about the translation from English to Spanish of an expression like "Friday the 13th".

(16) John stays home on Friday the 13th.

(17) John se queda en casa el viernes 13.
John stays home on Friday the 13th

(18) John se queda en casa el martes 13.
John stays home on Tuesday the 13th

The sentence (17) is the word for word translation of (16) and both the lexical items and the syntax of the sentence are equivalent. Traditional sentence similarity measures would thus give a value of 1. If we now consider both the original sentence and the implicated meaning, we arrive at a different result. From (16), we would naturally infer that the reason Luke stays home on this date is due to superstition because it is common knowledge in the anglo-american culture that this date is supposed to be unlucky. However, we could not derive this information from (17) because in some spanish-speaking countries, the usual unlucky day is Tuesday the 13th (as in (18)). In fact the cosine similarity value between the complete sentence-vectors of (16) and (18) is 0.9997 using the globalist view under SMCM while the similarity value between the complete sentences-vectors of (16) and of (17) is only 0.8843. This example shows that complete similarity, i.e. similarity of complete sentence-vectors, may play an important role in translation. It is important to note that for this particular example, we can only use the globalist view because it is not a scalar implicature and the localist view is not yet able to take into account such implicatures. The post-compositional structure of the globalist view allows for an easy integration of any kind of implicatures as long as the two contributions (*"What is said"* and *"What is implicated"* are already composed when combined together.

7 Conclusion

This paper was a first step in a new direction, bridging compositional distributional models and approaches to SI by looking at how implicated meaning could be treated within these models. Our goal was to emphasise the importance of the implicated meaning for complete sentence similarity measures and to acknowledge the discrepancies between different applications of the integration of SIs within CDS. We looked at two views regarding how the derivation of SIs takes place and we showed those two views were not treated the same way by two CDS models. This lead to the realisation that there are some differences in compositional treatments between the localist and the globalist view even when considering non-embedded sentences such as *John ate some candies*. We then derive the sentence-vectors and compute the difference in sentence similarity measures between the two views. We discussed the very low contribution from the implicated part of the sentence in the localist view and the inherent capacity for the globalist view to differentiate between the relative contribution of the explicit and the implicit meaning of a sentence and the fact that it can encompass more kinds of implicated meaning compared to the localist view [6].

References

1. G. Chierchia, D. Fox, and B. Spector. The Grammatical View of Scalar Implicatures and the Relationship between Semantics and Pragmatics. In P. Portner, C. Maienborn, and K. von Heusinger, editors, *Handbook of Semantics*. Mouton de Gruyter, 2011.
2. S. Clark. Vector Space Models of Lexical Meaning. In *The Handbook of Contemporary Semantic Theory*, pages 493–522. John Wiley & Sons, Ltd, Chichester, UK, aug 2015.
3. B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical Foundations for a Compositional Distributional Model of Meaning. *Linguistic Analysis*, 36(345), 2010.
4. F. Ervas. On Semantic and Pragmatic Equivalence in Translation. In B. by Pasa and L. Morra, editors, *Translating the DCFR and Drafting the CESL : A Pragmatic Perspective*. de Gruyter, 2014.
5. D. Fox. Free Choice and the Theory of Scalar Implicatures. *Presupposition and Implicature in Compositional Semantics*, pages 71–120, 2007.
6. B. Geurts. *Quantity Implicatures*. Cambridge University Press, 2010.
7. E. Grefenstette. *Category-Theoretic Quantitative Compositional Distributional Models of Natural Language Semantics*. PhD thesis, University of Oxford, 2013.
8. L. Horn. *On the Semantic Properties of Logical Operators in English*. PhD thesis, UCLA, Los Angeles, 1972.
9. D. Kartsaklis. Coordination in Categorical Compositional Distributional Semantics. *Electronic Proceedings in Theoretical Computer Science*, 221:29–38, 2016.
10. N. Koleva, A. Horbach, A. Palmer, and S. Ostermann. Paraphrase Detection for Short Answer Scoring. *NEALT Proceedings Series*, 22:59–73, 2014.
11. J. Lambek. From word to sentence. *Polimetrica, Milan*, 2008.
12. T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in Pre-Training Distributed Word Representations. *arXiv:1712.09405 [cs.CL]*, 2017.
13. J. Mitchell and M. Lapata. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429, 2010.

Towards an analysis of agent-oriented manner adverbials in German

Ekaterina Gabrovska

Heinrich Heine University, Universitätstr. 1, 40225 Düsseldorf, Germany
egabrovska@phil.uni-duesseldorf.de

Abstract. The paper discuses the status of the agent-relatedness of German agent-oriented manner adverbials, based on a case study of the adverbial *sorgfältig* ('carefully'). We claim that their treatment as agent-oriented is due to a mental-attitude-like component of their meaning, i.e. they specify the content of the intention of an agent. The combination of the mental-attitude and manner aspects in the meaning contribution of the modifier is explained based on Goldman's [7] theory of human action as it provides a tool which can capture the meaning components of the adverbial. The analysis is formalized in Frame Semantics as introduced in Petersen [13] and Löbner [10].

Keywords: lexical semantics · agent-relatedness · manner adverbials · frame semantics · actions.

1 Introduction

The investigation of adverbial modifiers sheds light on the semantics of events as well as agentivity. Especially interesting with respect to these two issues are agentive adverbials, such as agent-oriented and mental-attitude adverbials (cf. Ernst [5] a.o.), which have been extensively explored in the literature. Nevertheless, the exact nature of the agent-relatedness in connection to manner adverbials, like the class of German agent-oriented manner adverbials, e.g. *sorgfältig* 'carefully', *vorsichtig* 'cautiously', proposed in Schäfer [14], have been mostly left aside as far as their agent-relatedness is concerned. These adverbials, however, can provide us with new insights on the nature of manners of events as well as the role of the agent. Hence, what we aim for in this paper is to clarify the nature of the agent-relatedness of agent-oriented manner adverbials as well as the impact of this relation on the manner of the modified event. We approach this issue by analyzing the meaning contribution of such adverbials based on a case study of the modifier *sorgfältig* (although *sorgfältig* is translated here as *carefully*, the English adverbial has a broader meaning).

2 Proposal and Data

We propose that the modifier combines components characteristic of mental-attitude as well as manner adverbials. *Sorgfältig* ('carefully') demands the par-

ticipation of an intentional and controlling agent in the event and fixes the interpretation of potentially intentional verbs or applies to ones which can only be interpreted as intentional.[1] Furthermore, the modifier imposes certain restrictions on the agent of the modified action. We propose that these restrictions are best captured by Goldman's [7] 'action-plans' which are assumed to be at the base of an analysis of intentional action.

2.1 Goldman's treatment of (intentional) action

Goldman [7] offers a multi-level view on human action (cf. also Löbner [12]). Although the author takes as a starting point the philosophical discussion on act individuation, his theory is meant to be a cognitive one (cf. Goldman [8], Löbner [12]). Goldman [7], [8] assumes that the examples in (1) describe different acts. More precisely, the author considers the examples as describing one doing of an agent in the real world as a "combination of distinct act-tokens of distinct act-types" (Löbner [12]). The description is based on ways of categorizing actions, which are then seen as different levels in the conceptual representation (cf. Löbner [12]).

(1) (Example partly adopted from Goldman [7])
   a. John flips the switch.
   b. John turns on the light.
   c. John wakes up Mary.

According to Löbner's [12] discussion of Goldman [7], there is one doing of John out there in the real world, but this doing is conceptualized/categorized as a complex relation between acts of different types (cf. Löbner [12]). The acts are ordered in levels and related by the so-called 'level-generational' relation. This relation is a conceptual one: it applies at the "level of conceptual representation, or categorization, of actions" (Löbner [12]) and is based on the types of the involved act-tokens. The relation is furthermore asymmetric, irreflexive and transitive[2] and can be expressed by the use of the locution 'by' (or 'in'): *John turns on the light by flipping the switch* (cf. Goldman [7]). The level-generational relation captures that an act under suitable circumstances generates another act of the same agent at the same time (cf. also Löbner [12]).

The relation is actually quite intuitive: a waking up of Mary can be realized in many different ways. The verb itself does not specify a single method but depending on the circumstances of the action it could be realized by an act out of a set of possible methods (turning on the TV, touching her, calling her, and so on). Under some given circumstances, John uses the method of turning on the light to realize the waking up of Mary. If the circumstances of his action do not support the method, i.e. there is no bulb in the lamp, the act of flipping the

[1] A similar assumption is made for mental-attitude adverbials by Buscher [2].

[2] The generational relations, proposed by Goldman, capture different circumstantial connections between acts of different types (cf. Löbner [12]). Due to limitation of space, we refer the interested reader to Goldman [7].

switch do not have to enable the realization of the turning on of the light, i.e. it does not generate the higher act. Likewise, John's flipping the switch, under circumstances, generates (is a method of) turning on the light.

That the single acts in (1) are not identical is indicated by the causal relations between them (cf. Goldman [7] for a discussion of identity). The act of flipping the switch can cause the light to go on, but the act of turning on the light does not cause the flipping of the switch. Hence, the two acts do not have the same properties, which they should if considered identical according to Goldman [7].

Goldman ([7]: 57) further proposes that an agent is acting intentionally only if he has an action-plan which matches the realization of her action. This action-plan consists of an action-want, i.e. the want to realize a certain act, and the level-generational beliefs of the agent. The latter are understood as hypothetical acts of an agent related by level-generation.

For an intuitive illustration of action-plans consider an example by Goldman: *John turns on the light by flipping the switch.* If John is to turn on the light intentionally, then he has to have the want to turn it on. This want is part of his action-plan which causes his actual action (see Goldman [7]). Now, John can further have different options for turning the light on, but as he decides to use the switch, he believes that this method is going to realize the act he desires. Hence, in order to act intentionally, John must have an action-plan consisting of his want and a set of beliefs concerning the methods of realization of the target of that want.

We propose, based on Goldman, that the content of the intention of the agent is the target of her action-want. This intention then predetermines, according to the level-generational beliefs of the agent, how it could be realized.[3]

2.2 Analyzing *sorgfältig*

According to our proposal, the modifier *sorgfältig* needs an intentional and controlling agent to participate in the event. This means that an agent with an action-plan has to be involved. Furthermore, the modifier imposes restrictions on the manner of the event which we see as related to the content of the agent's intention.

The agent The assumption that an agent has to be present is already indicated by the name of the class. However, it has never been discussed in detail whether the involved participant is really to be seen as an agent. Similarly, mental-attitude adverbials are assumed to require an agent too, but as shown by Buscher [2], this is not always the case. Following her analysis, the subgroup of assimilative adverbials, like *widerwillig* ('reluctantly'), demand a controlling participant, whereas intentional adverbials, like *absichtlich* ('intentionally'), a participant initiating the event, but none of them demands a fully specified agent.[4] Thus, it

[3] The assumption also allows us to treat mental-attitude adverbials, as *absichtlich* ('intentionally'), as stating that an agent is acting with respect to her action-plan. A detailed analysis of *absichtlich* is going to be presented in Gabrovska et al. [6].

[4] See Buscher [2] for a discussion and supporting data.

has to be clarified first whether an agent (an intentional, volitional, and sentient being[5]) is involved. Consider the following examples:

(2) a. #*John hat sich sorgfältig verlaufen.* [6]
John has himself carefully get-lost
'John got lost carefully.'
b. #*John rutschte sorgfältig aus.*
John slipped carefully PART
'John slipped carefully.'

The examples in (2a) and (2b) show that *sorgfältig* cannot modify verbs denying control as well as unaccusative verbs (they do not assign an agent role (cf. Buscher [2]: 103ff among others)), respectively. Example (2) speaks in favor of the assumption that verbs modified by *sorgfältig* ('carefully') have to allow for an intentional interpretation (or have only an intentional interpretation, considering verbs like *arbeiten* ('to work') or *planen* ('to plan')).

(3) a. *Das Kind hat das Bild absichtlich/unabsichtlich*
The child has the picture intentionally/unintentionally
zerschnitten.
cut
'The child cut the picture intentionally/unintentionally.'
b. *Das Kind hat das Bild sorgfältig zerschnitten.*
The child has the picture carefully cut
'The child cut the picture carefully.'

Example (3a) shows that *zerschneiden* ('to cut') can be done intentionally or unintentionally. However, when the verb is modified by *sorgfältig*, (3b), only an intentional interpretation is possible. Even if we have a context where a child holds two pictures, one of his mother and one of his father, but is not aware that she holds them both as the picture of the mother is on top of the picture of the father hiding it completely. Now, if the child cuts *sorgfältig* the picture of her mother, it cannot be said to cut *sorgfältig* the picture of her father too.

Altogether, the data presented so far supports the assumption that an agent has to be present whenever the modifier *sorgfältig* ('carefully') is used. The example in (3) indicate that when combined with verbs which can be either intentional or unintentional, the modifier fixes the interpretation to the intentional variant.[7] Hence, the involved agent is acting intentionally. We use Goldman's action-plans to capture and handle intentional action. Hence, whenever an agent is acting intentionally, she is going to have an action-plan. This means that the agent has

[5] We follow Van Valin and Wilkins [16] as far as the notions *volition* and *sentience* are concerned.

[6] The # signals oddness due to meaning, whereas the * symbol means ungrammaticality.

[7] Buscher [2] states that mental-attitude adverbials, like *absichtlich* ('intentionally'), also fix the interpretation of verbs. We assume that this common feature between the two classes is to be explained by the use of the same mechanisms.

an action-want and chooses with respect to her beliefs the appropriate way of achieving, i.e. generating, the desired act.

The result The agent, if supposed to act *sorgfältig* ('carefully'), has to meet some further requirements concerning her intention and the way her beliefs would realize that intention.

The notion of 'result' as used here refers to the target of the action-want of the agent, i.e. the desired act, not a produced object or a resulting event. Hence, the result is the target of the agent's intention.

(4) a. *Wir haben sorgfältig gearbeitet, weil wir gute Ergebnisse*
We have carefully worked, because we good results
erzielen wollten.
achieve wanted
'We worked carefully, because we wanted to achieve good results.'

b. #*Wir haben sorgfältig gearbeitet, weil wir schlechte Ergebnisse*
We have carefully worked, because we bad results
erzielen wollten.
achieve wanted
'We worked carefully, because we wanted to achieve bad results.'

Both examples verbalize the agent's desire: achieving good/bad results. The difference in acceptability between (4a) and (4b) is due to the quality of the result. Under normal circumstances a *sorgfältig* ('careful') action is unlikely to be performed if the quality of the result is intended to be low. Assuming a context where the agent actually desires to achieve bad results for some reason, the sentence in (4b) can be felicitous. The observation is further supported by the following data:

(5) a. *Wir haben sorgfältig gearbeitet, daher haben wir gute*
We have carefully worked, therefore have we good
Ergebnisse.
results
'We worked carefully, therefore our results are good.'

b. #*Wir haben sorgfältig gearbeitet, daher haben wir schlechte*
We have carefully worked, therefore have we bad
Ergebnisse.
results
'We worked carefully, therefore our results are bad.'

Comparing (5a) and (5b) we see that it is the quality of the achieved results which indicates the oddness of the latter. Data like this shows that a *sorgfältig* action is more probable to lead to a "good" result.[8] Although the example in

[8] It has to be noted that *gute Ergebnisse* ('good results') or *having good results* is not an act as required by Goldman ([7]: 52f). In accordance with the author, the corresponding desired act is to achieve/gain good results.

(5b), as (4b), is not completely uninterpretable, additional context is necessary to make the sentence acceptable, whereas (5a), like (4a), is fine on its own.

This specification of the result as one with a high quality is, however, a context-dependent inference as indicated by the next example.

(6) *John hat sein Zimmer sorgfältig geputzt, trotzdem hingen*
John has his room carefully cleaned nevertheless hung
Spinnweben an der Decke.
spider webs on the ceiling
'John cleaned his room carefully but nevertheless spider webs hung from the ceiling.'

Although the cleaning is said to be *sorgfältig*, the achieved result is not seen as "good" enough. Such data indicates that the target of the agent's want does not have to be realized. Nevertheless, the cleaning itself is still intentional and its method of implementation still suitable for the achieving of at least part of the relevant aspects of the intended result. Furthermore, native speakers, when confronted with such data, tend to assume reasons outside of the control of the agent as an excuse. Hence, in the example above, the agent is assumed to be unable to clean the spider web due to the height of the ceiling, for example. This leads to the assumption that the quality of the result is measured with respect to the abilities and control possibilities of the agent.

The method The inference of a "good" result evoked by *sorgfältig* ('carefully') also depends on the manner in which the action is realized.

We call this manner component the method of realization of the desired act and propose that abstract actions have more than one possible method of realization. This assumption is in line with Löbner's [12] view on the lexical meaning of action verbs, where Goldman's perspective is adopted for concepts constituting the lexical meanings of the latter.

Considering again Goldman's [7] theory, the relation between acts is phrased by the use of 'by' (translated as *indem* in German[9]). Hence, in the example *John turns on the light by flipping the switch by moving his hand*, 'by' states that the act of flipping the switch is the method of realizing the act of turning on the light. This means that the act of flipping the switch is suitable for generating the act of turning on the light under the given circumstances. Similarly, all more abstract actions have at least one or more methods of realization.

(7) *Die Rüben werden sorgfältig gereinigt, indem man sie einige*
The turnips shall carefully cleaned, by one them several
Minuten lang im Wasser läßt, dann wäscht und abbürstet.[10]
minutes long in.DAT water leaves, then washes and brushes down

[9] Not every use of *indem* ('by') signals a level-generational relation. Mere instruments can also be introduced by the connector. For an analysis of *indem* see cf. Bücking [3].

'The turnips are cleaned carefully, by leaving them in water for several minutes and then washing and brushing them down.'

Example (7) supplies the method suitable of realizing a *sorgfältig* cleaning. The method is complex and consists of the sequential realization of three acts: putting in water, washing, and brushing down. If one of these acts is left out, the cleaning is no longer *sorgfältig*. Hence, not all methods implementing a cleaning are available when an action is said to be *sorgfältig* ('carefully'). This is not unexpected, considering that the result of the action should have a rather high quality. It is only natural that if a "good" result is expected, not every method can guarantee high quality. Nevertheless, the number of possible methods is usually not reduced to exactly one by the adverbial. Rather, the set of methods implementing a certain action is specified more concretely and therefore reduced in the number of elements compared to the set containing all possible methods. Here again, the method and its suitability depend on the context in which the action is performed. Existing (social) conventions and rules also have an impact on the suitability.

3 The analysis

Bringing all the parts together we can state that *sorgfältig* ('carefully') demands the participation of a controlling agent with the intention for a result with a high quality. The modifier further relates this intention to the method suitable for the realization. As the agent is acting intentionally, she is the one choosing which method is suitable for the realization of her want with respect to her abilities as well as the circumstances and conventions/standards holding at the moment of realization and her knowledge of these.

3.1 The formalization

The proposal, as presented so far, is formalized in the framework of Düsseldorf Frame Semantics, which adopts and develops Barsalou's [1] proposal on the representation of knowledge in human cognition. The theory is based on the Frame Hypothesis, which states that frames are the general format of representation in human cognition (Löbner [9]).

In this framework, frames are defined as recursive attribute-value structures, where attributes are functional and are restricted to only one sort[11] (Petersen

[10] https://books.google.de/books?id=FXGXBwAAQBAJ&pg=PA133&lpg=PA133&dq=%22sorgf%C3%A4ltig+gereinigt,+indem%22&source=bl&ots=e_6WSP7pZ1&sig=JT1a1IywRLv4QFJTUqW7eapOwRU&hl=de&sa=X&ved=0ahUKEwjo-KPH5o7ZAhXD1qQKHd3ACwQQ6AEIJzAA, last accessed 27.06.18, 11:52.

[11] Sorts are seen as maximal types in a hierarchy of types (Löbner [10]). Löbner [10] considers sorts an a priori part of his frame ontology, whereas types are derived from sorts and attributes, e.g. 'color' is a sort, whereas 'blue' is a type.

[13], Löbner [9], [10], [11]). They can be represented as diagrams, attribute-value matrices or in predicate logic. The manner of representation adopted here are diagrams. Hence, a frame consists of nodes connected by arcs. The nodes represent values, whereas arcs stand for attributes. As a frame is seen as a description of an individual, there is a central node marked with a double line standing for this individual. Furthermore, values of nodes are seen as belonging to a type which itself is part of a bigger type-hierarchy.

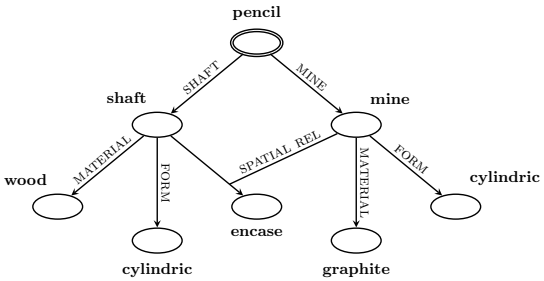


Fig. 1. Partial frame for 'pencil' based on Löbner ([11]:4, Figure 1)

The frame in fig. 1 represents the standard kind of pencil with a wooden shaft and a graphite mine. Thus, the central node of the frame is of type **pencil**.[12] The pencil consists of a shaft and a mine, represented by the attributes SHAFT and MINE. The mine and the shaft nodes are connected with each other by the two place attribute SPATIAL RELATION. As frames are recursive structures, the latter two nodes themselves can be represented as frames and have attributes as MATERIAL and FORM.

Goldman's [7] considerations on human action are captured in frames by the notion of *cascades*. Values of the nodes in a cascade are not individuals but first-order frames describing acts (i.e. events) of different types; cascades are second-order frame structures. The act-frames are related via a 'c-constitution' relation which is based on Goldman's 'level-generation'.[13] The c-constitution relation captures the fact that under circumstances an act can generate another act of a different type.

The cascade in fig. 2 represents that under given circumstances (e.g. Mary is sensitive to light, John knows that and there is a bulb in the lamp a.o.), the act of *flipping the switch* by John constitutes the act of *turning on the light* by the same agent. Likewise, the latter act constitutes the act of *waking up Mary*.

[12] Node types are marked by bold font, whereas attributes by small caps.

[13] See Löbner [12] for a definition of 'c-constitution' as well as for a discussion of 'level-generation'.

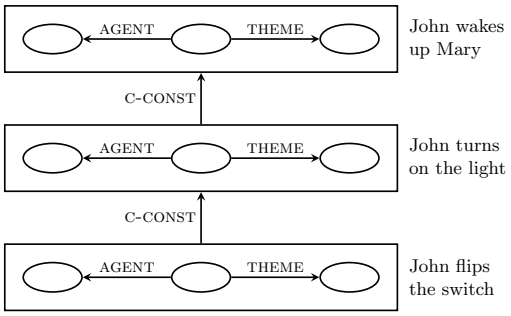


Fig. 2. Cascade

3.2 A model for *sorgfältig*

A frame representation capable of capturing the meaning contribution of *sorgfältig* ('carefully') has to offer access to the intention of the agent as well as to the action realizing that intention. Hence, both a plan and an action-realization representation have to be available.

For this purpose we treat actions as involving a plan and a realization representation. This kind of representation is inspired by the analysis of causation events in Kallmeyer & Osswald [15], where the standard template-based event structure for causative verbs[14], e.g. *break*, is represented in frames. Hence, the frame in fig. 3 has as a central node an action with the attributes PLAN and EXECUTION which assign an action-plan, more precisely a level-generational beliefs construction including the target of the agent's want, and a physical realization of an action respectively. As already discussed an action-plan as well as an action are represented themselves as cascades. Thus, the values of the plan and the execution nodes are cascades.

Following Goldman [7], an action is intentional if realized as conceived in the agent's action-plan. Hence, the plan and the action have to match. This means that values of the action cascade nodes have to match the respective values of the action-plan cascade. This is achieved by the use of a comparator[15], which checks whether the acts at a certain level have the same type or not. If they do then the values match, otherwise not.

[14] While Kallmeyer & Osswald [15] concentrate on the primitive CAUSE, we attend to Dowty's [4] DO. The difference here is that we do not assume the agent to be an argument of DO, but rather the agent's action-plan (more details concerning the formalization are going to be presented in Gabrovska et al. [6]).

[15] Comparators are defined in Löbner [10] as two-place attributes with arguments of the same sort which return comparison values, for example '=', '>', and '<' (not to be confused with the relations these symbols as taken to denote). These attributes are used for the representation of intrasortal relations (see Löbner [10] for more detail).

The matching, however, does not have to occur for all levels of the cascade. The result does not have to be actually achieved, but only desired. If the method intended to implement the result is suitable and realized as conceived in the plan, the agent's action can be categorized as *sorgfältig*. It is, nevertheless, necessary that the agent desires a result with a certain quality.

Sorgfältig further postulates restrictions on the target of the agent's want and the method of realization of that target. These restrictions are captured by value constraints.[16] The values of the highest node of the action-plan cascade are constrained to high quality results. Such results cannot be intentionally achieved by doing just anything, hence the restriction of result values leads to restrictions on the method values, the lower nodes in the action-plan cascade. Constraints, however, are not visible in a frame diagram.

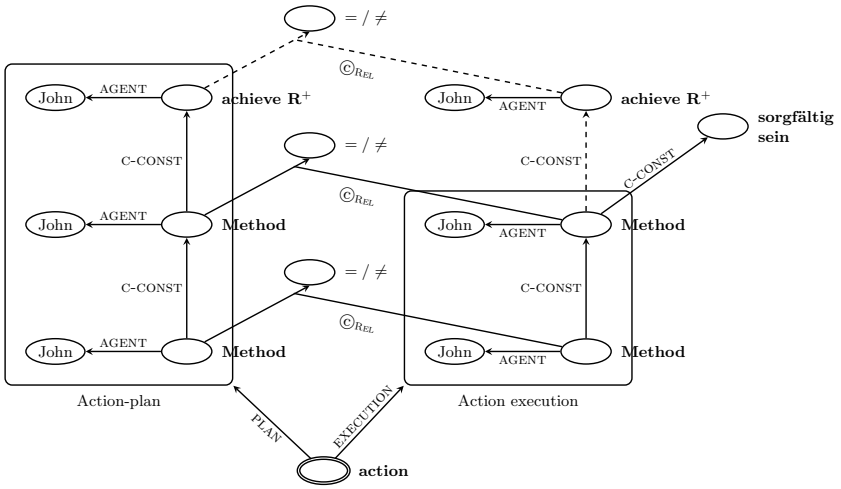


Fig. 3. Connecting action and action-plan

Fig. 3 shows an example representation of a *sorgfältig* action. The agent is acting intentionally, which is represented by the fact that **Method** on the right hand side is congruent with the plan on the left. The restrictions the modifier makes on the target of the agent's want and the method of implementation are constraints between the values of the two nodes connected by the C-CONST relation in the action-plan cascade. As the methods of result implementation match the realization, the restrictions imposed by the modifier are realized. This means that the method is such that it can generate a *sorgfältig sein* ('being careful') level in the execution track of the frame. The *sorgfältig sein* node is

[16] Constraints are, following Barsalou [1], a means capturing the interactions of values of attributes.

a further categorization of the agent's act (cf. Dowty [4]:114f on the agentive reading of careful). The result component, if realized, is also generated by the method and matches the intended result. The possibility that the result might not be achieved is marked in the frame by the use of dashed lines.

Let us illustrate the analysis on an example.

(8) *Ich habe mein Zimmer sorgfältig geputzt, indem ich gesaugt habe*
I have my room carefully cleaned by I vacuumed have
und alle Spinnweben entfernte.
and spider webs removed
'I cleaned my room carefully by vacuuming and removing all spider webs.'

Assume that John is the agent, who is cleaning *sorgfältig* according to his own criteria and knowledge of the relevant conventions and rules. The method of realization is complex and consists of *saugen* ('to vacuum') and *Spinnweben entfernen* ('to remove spider webs'). John's desired act (the result to be achieved) is to achieve that his room is well cleaned (clean^+) and according to his beliefs the chosen method can generate this act. The lowest level in fig. 3 stands for the vacuuming and removing spider webs; the intermediate level represents the cleaning, which is implemented by the vacuuming and removing spider webs; on top, generated by the cleaning is the result *achieve having well cleaned room* (as the result is achieved in this case the dashed lines should be normal lines). The cleaning is further categorized as *sorgfältig sein*. As in this case plan and execution match the comparator returns the value '=' for each level.

4 Summary and Outlook

So far we have assumed that intentional agents have action-plans as proposed by Goldman (1970) and that *sorgfältig* ('carefully') relates the desire and the belief components of the action-plan in a specific way, stating that the method chosen for the realization of the desire has to be suitable with respect to the intended quality of the result. Hence, as a mental-attitude adverbial, *sorgfältig* attributes an intention to the agent and at the same time, as a manner adverbial, specifies how the action should be realized with respect to this intention. The formalization of the proposal is realized in the framework of frame semantics as in Petersen [13] and Löbner [10] [12] and is capable of capturing agent-relatedness as well as its impact on the manner components of events.

The tentative analysis presented here still leaves a number of open questions for future work. But is has the potential of explaining the meaning contribution of not only *sorgfältig* ('carefully'). We assume that agent-oriented manner adverbials assign an intention to the participating agent and relate this intention to the method of realization. The difference between modifiers of this class is then the content of the intention and its impact on the method of realization. Furthermore, the meaning contribution of mental-attitude adverbials, like *absichtlich* ('intentionally'), can also be captured by the assumption of action-plans.

Acknowledgements

Many thanks to Sebastian Löbner, Willi Geuder, Curt Anderson, the anonymous reviewers as well as all my informants for the long discussions, the comments, the data, and all other kinds of help. The work is supported by DFG CRC 991 "The Structure of Representations in Language, Cognition, and Science," project B09.

References

1. Barsalou, L. W. Frames, concepts, and conceptual fields. In A. Lehrer, E. F. Kittay (Eds.), Frames, fields, and contrasts: New essays in semantic and lexical organization, pp. 21–74. Hillsdale, NJ: Lawrence Erlbaum Associates (1992)
2. Buscher. F. Kompositionalität und ihre Freiräume: Zur flexiblen Interpretation von Einstellungsadverbialen. Dissertation, Universität Tübingen (2016)
3. Bücking, S. Elaborating on Events by means of English by and German indem. In C. Piñón (ed.), Empirical Issues in Syntax and Semantics 10, 19–36, www.cssp.cnrs.fr/eiss10/, (2014)
4. Dowty, D. R. Word Meaning and Montague Grammar. Dordrecht/Boston/London: D. Reidel Publishing Company (1979)
5. Ernst, T. The Syntax of Adjuncts. Cambridge: Cambridge University Press (2002)
6. Gabrovska, E & W. Geuder & S. Löbner. The Conceptual Structure of Action and Its Modifiers (in prep.)
7. Goldman, A. I. A theory of human action. New Jersey: Prentice-Hall, INC (1970)
8. Goldman, A. I. Action, causation, and unity. Noûs, 13, 261–270 (1979)
9. Löbner, S. Evidence for frames from human language. In T. Gamerschlag, D. Gerland, R. Osswald, W. Petersen (Eds.), Frames and concept types, pp. 23–67. Heidelberg, New York: Springer (2014)
10. Löbner, S. Frame theory with first-order comparators: modeling the lexical meaning of punctual verbs of change with frames. In H. H. Hansen, S. E. Murray, M. Sadrzadeh, H. Zeevat (Eds.), Logic, Language, and Computa-tion. 11th International Tbilisi Symposium, pp. 98–117. Heidelberg, New York: Springer (2017)
11. Löbner, S. 'Barsalou-Frames in Wort- und Satzsemantik'. In Jahrbuch 2017 des Instituts für Deutsche Sprache. Berlin, Boston: De Gruyter (to appear 2018)
12. Löbner, S. Cascades. Goldman's level-generation, multilevel concept of action, and verb semantics. manuscript. http://www.sfb991.uni-duesseldorf.de/fileadmin/Vhosts/SFB991/b09/Loebner_Cascades_ms_04_18.pdf
13. Petersen, W. 2007. Representation of Concepts as Frames. In: Complex Cognition and Qualitative Science, Jurgis Skilters, Fiorenza Toccafondi and Gerhard Stemberger (eds.), The Baltic International Yearbook of Cognition, Logic and Communication, 2, p. 151–170. University of Latvia.
14. Schäfer, M. 2013. Positions and interpretations. German adverbial adjectives at the syntax-semantics interface. Berlin: De Gruyter Mouton.
15. Kallmeyer, Laura; Osswald, Rainer (2013): Syntax-driven semantic frame composition in lexicalized tree adjoining grammars. In Journal of Language Modelling 1 (2), pp. 267–330.
16. Van Valin, R. D. Jr. & D. P. Wilkins. The case for 'effector': Case roles, agents, and agency revisited. In M. Shibatani & S. A. Thompson (eds.), Grammatical constructions, 289–322. Oxford University Press. (1996)

Metafictional anaphora: A comparison of different accounts

Merel Semeijn

University of Groningen, Oude Boteringestraat 52, The Netherlands
m.semeijn@rug.nl
https://merelsemeijn.wordpress.com/

Abstract. I argue that pronominal anaphora across mixed parafictional/metafictional discourse (e.g. *In* The Lord of the Rings*, Frodo$_i$ goes through an immense mental struggle. He$_i$ is an intriguing fictional character!*) poses a problem for a workspace account. I evaluate different possible solutions based on a descriptivist approach, Zalta's logic of abstract objects and Recanati's dot-object theory.

Keywords: Metafictional statement · Workspace account · Anaphora · descriptivist approach · Abstract objects · Dot-objects.

Semanticists of fiction typically distinguish between (at least) three different kinds of statements that contain fictional names (e.g. 'Frodo'): In Recanati's [17] terminology, 'fictional', 'parafictional' and 'metafictional' statements. Fictional statements are statements taken directly from some fictional work (e.g. (1) from *The Lord of the Rings*). Parafictional statements are statements about the content of a fictional work (e.g. (2) or (3) as found in a discussion on *The Lord of the Rings*) that can be either 'explicit' (2) or 'implicit' (3), depending on whether the 'In fiction x,'-prefix is overt or covert. Metafictional statements are statements about fictional entities *as fictional entities* (e.g. (4)):

(1) Frodo had a very trying time that afternoon.

(2) In *The Lord of the Rings*, Frodo is a hobbit living in the Shire.

(3) Frodo is a hobbit living in the Shire.

(4) Frodo is an intriguing fictional character.

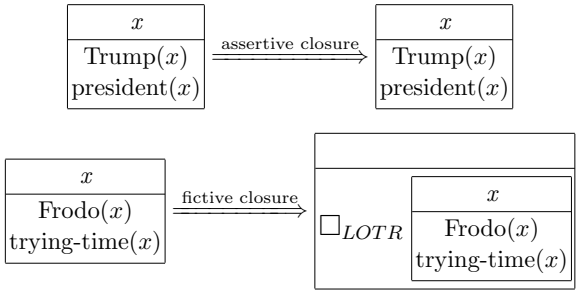
'The workspace account'[1] is a Stalnakerian approach to modelling fictional and parafictional statements. In this paper I briefly introduce the workspace account (Section 1) and argue that pronominal anaphora across mixed parafictional/metafictional discourse (henceforth 'metafictional anaphora') poses a challenge for the workspace account (Section 2). I explore and evaluate three different possible solutions based respectively on a descriptivist analysis of pronouns (Section 3.1), Zalta's abstract object theory [22, 23] (Section 3.2) and Recanati's dot-object theory [17] (Section 3.3).

[1] For details, see Semeijn [18].

1 Introducing the workspace account

In Stalnaker's [19] widely adopted pragmatic framework, assertions are modelled as proposals to update the 'common ground' (i.e. the set of mutually presupposed propositions between speaker and addressee). Previous attempts to extend the Stalnakerian framework to fiction (Stokke [20] and Eckardt [3]) are compatible with the consensus view of fiction interpretation (e.g. Currie [2] or Walton [21]) that links fiction to the cognitive attitude of imagination (i.e. regular assertions are mandates to *believe* and fictional statements are mandates to *imagine*). Likewise, fictional statements are modelled as proposals to update an 'unofficial' common ground that is separate from the 'official' common ground.

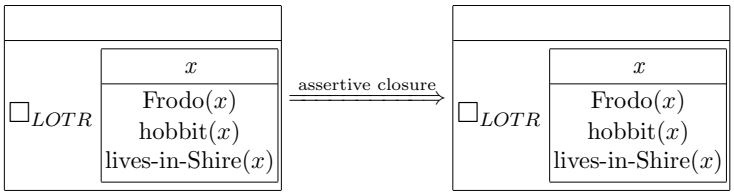
The workspace account is compatible with Matravers' [15] theory of fiction. He argues against the consensus view that there is no special cognitive attitude involved in our engagement with fiction. In fact, our primary engagement with narratives – whether fictional or non-fictional – involves the same cognitive processes. Likewise, in the workspace account, fictional statements and regular assertions are modelled as proposals to update the same *temporal* common ground: the workspace. What differentiates non-fiction from fiction is whether, at the end of the possibly multi-sentence discourse, 'assertive' or 'fictive closure' is performed; Whether the content of the updated workspace is added to the (official) common ground as belief (for non-fiction) or as parafictional belief (for fiction) under the relevant fiction-operator. I present a simplified representation of assertive closure of the assertion *Trump is the president of the U.S.* and of fictive closure of (1). To represent the workspace and common ground, I use the box notation of DRT (Discourse Representation Theory) developed by Kamp [9] in which NP's in a discourse are mapped to 'discourse referents' placed under several conditions:



Whether I am reading a fictional narrative (e.g. *The Lord of the Rings*) or a non-fictional narrative (e.g. some article in *The Times*), I update the workspace with the content of the narrative. As soon as I stop entertaining the propositions of a non-fictional narrative I perform assertive closure: I stop updating the workspace (with e.g. *Trump is the president of the U.S.*), and instead update the common ground with this information (i.e. I adopt it as belief). As soon as I stop entertaining the propositions of a fictional narrative I perform fictive closure: I stop updating the workspace with fictional statements (e.g. *Frodo had*

a very trying time on a particular afternoon) and instead update the common ground with *parafictional statements* based on the content of the workspace (e.g. *In* The Lord of the Rings*, Frodo had a very trying time on a particular afternoon*). Hence, after engaging with *The Lord of the Rings* all that I am left with are parafictional beliefs about its content.

Parafictional statements such as (2) are modelled as regular assertions about the content of a particular novel. Hence after engaging in parafictional discourse we perform assertive closure; The content of the workspace is added directly to the common ground as belief:



Thus, you end up with parafictional beliefs both after engaging reading *The Lord of the Rings* and after engaging in a conversation about its content. What differs in these two cases is the content of the workspace (i.e. whether you entertained propositions about hobbits or about the content of a particular novel).

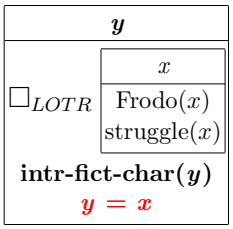
2 The challenge of metafictional anaphora

Metafictional statements are not modelled in the workspace account and pose a problem. More specifically, pronominal anaphora across mixed parafictional/metafictional discourse pose a problem. Consider the following felicitous discourse:

(5) In *The Lord of the Rings*, Frodo$_i$ goes through an immense mental struggle to save his$_i$ friends. Ah yes, he$_i$ is an intriguing fictional character![2]

The last sentence in (5) is a metafictional statement because it is about Frodo *as a fictional entity*. As Lewis [12] argued, metafictional statements are not covertly embedded under fiction-operators. For instance, the metafictional statement in (5) does not express that in *The Lord of the Rings*, Frodo is an intriguing fictional character. Instead, it is a proposal to directly update the common ground (in bold) or, in other words, to perform assertive closure:

[2] For simplicity, I henceforth omit the anaphoric links of the possessive 'his'.



The metafictional statement in (5) contains a pronoun 'he' that is anaphoric on the name 'Frodo' introduced in the preceding *parafictional* statement. Standardly, we take this to mean that the two terms co-refer because of the so-called 'Anaphora-Coreference Principle' (i.e. if a pronoun is anaphoric on a antecedent name, the two terms co-refer). We represent this by equating their discourse referents (i.e. adding $x = y$ to the common ground). However, following standard DRT-rules, x is not accessible outside of the $LOTR$ fiction-operator. Hence, it is unclear how we can interpret metafictional statements such as the one in (5).

3 Comparison of different solutions

There are different strategies available to meet the described challenge. In this section I discuss and evaluate a descriptivist approach (Section 3.1), an abstract object account (Section 3.2) and Recanati's dot-object account (Section 3.3).

3.1 A descriptivist approach: A description of Frodo

A possible solution to the described challenge in a traditional semantics framework is a descriptivist approach to anaphora (e.g. Evans [5], Elbourne [4] or Heim [8]). This analysis was originally proposed to account for donkey anaphora without relying on a dynamic semantic approach. Consider the following donkey sentence (6):

(6) If Sarah owns a donkey, she beats it.

Intuitively, the pronoun 'it' does not refer to a particular individual donkey but is bounded by 'a donkey'. However, it is outside of the syntactic scope of 'a donkey'. On a descriptivist analysis, the anaphoric pronoun 'it' functions like, or 'goes proxy for', the definite description 'the donkey' retrieved from the preceding clause. In Elbourne's D-type account, this is because NPs at the level of syntax undergo phonetic deletion (are not pronounced at the surface level) when in the environment of an identical NP (e.g. *My shirt is the same as his*). Similarly, (6) is in fact equivalent to (7):

(7) If Sarah owns a donkey, she beats the donkey.

This analysis evades the problem of the unbindable pronoun by replacing it with a definite description.

When we apply this strategy to (5), the pronoun 'he' is also analysed as going proxy for a definite description retrieved from the previous clause. However, (5) cannot be the result of simple phonetic deletion of an identical NP. If it were, (5) would be equivalent to something like (8):

(8) In *The Lord of the Rings*, Frodo goes through an immense mental struggle to save his friends. Ah yes, the person named Frodo in *The Lord of the Rings* that goes through an immense mental struggle to save his friends, is an intriguing fictional character!

This gives us an incorrect analysis of (5): A flesh and blood person cannot be a fictional character. Rather, the required definite description is a *metafictional* description such as 'the *character* named Frodo in *The Lord of the Rings*' so that (5) becomes equivalent to:

(9) In *The Lord of the Rings*, Frodo goes through an immense mental struggle to save his friends. Ah yes, the character named Frodo in *The Lord of the Rings* is an intriguing fictional character!

Although, (9) gives an acceptable analysis of what is expressed by (5) it is unclear how to obtain such a meta-description of Frodo from the preceding clause. Moreover, even if we assume that we can accommodate such a definite description for metafictional anaphora, this solution does not extend to other types of mixed discourse such as pronominal anaphora across mixed metafictional/parafictional discourse (e.g. *Frodo$_i$ is an intriguing fictional character. Ah yes, in* The Lord of the Rings *he$_i$ goes through an immense mental struggle to save his friends!*) which would require accommodation of yet another type of definite description. Hence, a descriptivist approach does not (as yet) adequately account for metafictional anaphora; Simple phonetic deletion does not provide appropriate definite descriptions and hence we need an account of how to accommodate these.

3.2 Abstract object theory: Frodo the abstract object

An alternative strategy is to claim that fictional names in parafictional and metafictional statements refer to an object that is accessible in the main box. For example, in applying his logic of abstract objects [22, 23] to fiction, Zalta claims that parafictional and metafictional statements are about abstract objects (i.e. Frodo the fictional character) that really exist.

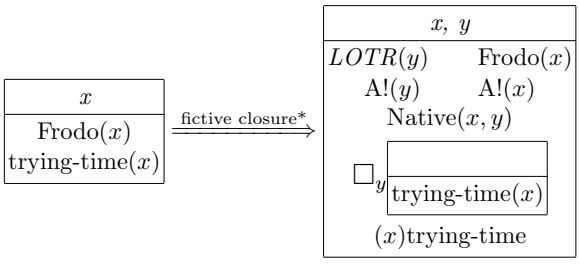
Zalta distinguishes two types of objects: x is an 'ordinary object' ('O!(x)') if it is, or could have been, concrete (e.g. a chair). x is an 'abstract object' ('A!(x)') just in case it could not be concrete (e.g. the empty set). There are two distinct kinds of predication: An ordinary object like a chair can 'exemplify' being red, i.e. it has the property of redness. Zalta denotes this using standard predicate logic notation: 'red(c)'. An abstract object can also 'encode' properties which means it has this property as one of its constitutive characteristics. For instance, the empty set encodes the property of having no members. This is denoted with the argument to the left of the predicate: '($\varnothing$)memberless'. Ordinary objects

do not encode properties but abstract objects do exemplify properties (e.g. the empty set exemplifies being well-discussed: 'well-disc($\varnothing$)').

A 'story' (e.g. *The Lord of the Rings*) is an abstract object that encodes the content of a narrative; It encodes 'vacuous properties' or propositional properties of the form 'being such that P is true', where P is a proposition that is true in the story. A 'fictional character' is an abstract object that is *native* to a story (e.g. Frodo or the One Ring, but not Napoleon).

Contrary to common practice, Zalta draws a strong distinction between explicit and implicit parafictional statements (e.g. respectively (2) and (3)). This is because Zalta is a realist about fictional characters (i.e. they exist as abstract objects) and hence we can talk about them as we do about ordinary objects (i.e. without an 'In fiction x'-operator or some type of pretense). A statement such as (3) is thus actually not 'implicit' in the sense that is has a covert fiction operator. Rather, it is a plain statement about what properties a certain abstract object encodes: '(f)hobbit $\wedge$ (f)lives-in-Shire'. Explicit parafictional statements (e.g. (2)) on the other hand do contain an 'In fiction x'-operator. They are statements about specific encoding and exemplifying relations between stories and characters. For instance (2) expresses that *The Lord of the Rings* encodes the property of being such that Frodo exemplifies being a hobbit that lives in the Shire: '$\Box_{LOTR}$hobbit(f)'. Metafictional statements are statements about what properties fictional characters *exemplify*. For instance, the metafictional statement in (5) expresses that Frodo exemplifies the property of being an intriguing fictional character: 'intr-fict-char(f)'.

Incorporating these ideas into the workspace account suggests a modification of the fictive closure operation. Because of the strong distinction drawn between the analysis of implicit and explicit parafictional statements, fictive closure can in theory involve two different kinds of updates of the common ground. I present a (simplified) representation of fictive closure* of (1) that includes updates of the common ground with both types of statements:[3]

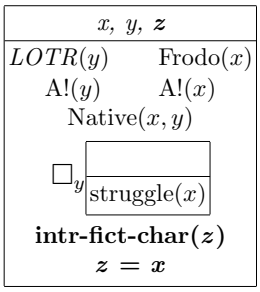


As soon as I stop reading *The Lord of the Rings*, I update the common ground with discourse referents for the newly introduced abstract objects (e.g. the fictional character Frodo) and with (explicit and implicit) parafictional beliefs based on the content of the workspace (e.g. '$\Box_{LOTR}$trying-time(f)' and

[3] Zalta adds a theorem to his theory ('$(x)(s)(Native(x,s) \rightarrow (F)(xF \equiv \Box_s Fx))$') according to which, if some character x is native to some story s, implicit and explicit parafictional statements about x necessarily follow from one another.

'(f)trying-time'). Importantly, not all propositional content of the workspace is updated as parafictional belief simpliciter; Proper name conditions (e.g. 'Frodo(x)') are separated from the other conditions in the workspace and placed in the main box.[4] This represents the fact that the abstract object Frodo is also named 'Frodo' outside of *The Lord of the Rings*.

In other words, I add an abstract object to the shared ontology for any fictional entity that is introduced and is native to the relevant story. This means that I incorporate Zalta's metaphysical assumptions that entail the existence of abstract objects in the actual world. It also means that after reading *The Lord of the Rings* the discourse referent for (the abstract object) 'Frodo' is accessible outside of the fiction-operator. This solves the challenge posed by metafictional anaphora. To see how we have to first recognize that because Zalta draws a strong distinction between implicit and explicit parafictional statements, the challenge splits up in two sub-challenges: One of pronominal anaphora across mixed *explicit* parafictional/metafictional discourse and one of pronominal anaphora across mixed *implicit* parafictional/metafictional discourse. Our central example up to this point (5) is an example of pronominal anaphora across mixed *explicit* parafictional/metafictional discourse. I represent the common ground based on the parafictional statement in (5) and the proposed metafictional update (in bold) as follows:



Next, we can rewrite (5) so that it is an example of pronominal anaphora across mixed *implicit* parafictional/metafictional discourse:

(10) Frodo$_i$ goes through an immense mental struggle to save his$_i$ friends. Ah yes, he$_i$ is an intriguing fictional character!

I represent the common ground based on the parafictional statement in (10) and the proposed metafictional update (in bold) as follows:

[4] Alternatively, we can model this as a doubling of the proper name condition so that it features both inside and outside the $LOTR$ fiction-operator.

x, y, $\boldsymbol{z}$	
$LOTR(y)$	Frodo(x)
A!(y)	A!(x)
Native(x, y)	
(x)struggle	
intr-fict-char($\boldsymbol{z}$)	
$\boldsymbol{z = x}$	

As the formalisms show, in both cases the discourse referent x for 'Frodo' is accessible outside of the $LOTR$ fiction-operator. Hence we can equate the discourse referents for 'Frodo' and 'he' and interpret the metafictional statements in (5) and (10).

Although this analysis seems to straightforwardly solve the problem of metafictional anaphora, on closer inspection Zalta's analysis of explicit parafictional statements is problematic. Remember that (2) expresses that the abstract object *The Lord of the Rings* encodes the vacuous property of being such that P (where P is the proposition that Frodo exemplifies being a hobbit that lives in the Shire). P is supposedly true according to *The Lord of the Rings*. However, the name 'Frodo' refers to an abstract object and hence it is true according to *The Lord of the Rings* that *the abstract object* Frodo exemplifies being a hobbit that lives in the Shire. This seems problematic; First, how can an abstract object *exemplify* being a hobbit or living in the Shire? These are the kind of properties that abstract objects *encode*. Moreover, intuitively *The Lord of the Rings* is a story about flesh and blood hobbits, not a story about what properties certain abstract objects exemplify or encode.[5] Hence, any analysis according to which fictional names that occur under a fiction-operator refer to abstract objects, seems problematic. Therefore, although an abstract object account solves the sub-challenge of pronominal anaphora across mixed *implicit* parafictional/metafictional discourse, it runs into difficulties with the sub-challenge of pronominal anaphora across mixed *explicit* parafictional/metafictional discourse. The only way to solve this problem seems to be to allow for some kind of ambiguity in the name 'Frodo' so that it refers to a flesh and blood individual when it occurs in the fiction operator and to an abstract object when it occurs outside of the fiction operator. This strategy is explored in the next section.

3.3 Dot-object theory: The different facets of Frodo

A different available strategy to solve the problem of metafictional anaphora is to claim that parafictional and metafictional statements are about different kinds of objects (e.g. Currie [2] or Kripke [11]), i.e. we do not add $y = x$ to the common ground when interpreting (5) and hence there is no accessibility problem.

Prima facie, the admissability of pronominal anaphora across mixed parafictional/metafictional discourse forms a problem for such an account because of the Anaphora-Coreference Principle. However, Recanati ([17]) argues there are apparent counterexamples to this principle. Take the following sentence:

[5] A similar concern has been voiced by Klauk [10].

(11) Lunch$_i$ was delicious, but it$_i$ took forever (adapted from Asher [1, p.11])

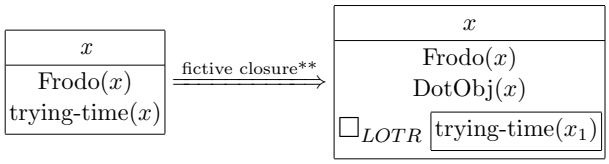
The pronoun 'it' is anaphoric on the noun 'lunch' of the preceding clause. However, 'lunch' and 'it' do not co-refer; 'lunch' refers to food (which was delicious) and 'it' refers to a social event (which took forever). Following Recanati, we can save the Anaphora-Coreference Principle by appealing to the notion of a so-called 'dot-object' (See e.g. Pustejovsky [16], Luo [13] or Asher [1]), i.e. "a complex entity involving several 'facets' " [17, p.15]. The noun 'lunch' is polysemous (i.e. it can refer to food or a social event) and hence denotes a dot-object (represented as `food • social event`) involving several facets (i.e. a food facet and a social event facet). Thus, in (11) 'lunch' and 'it' do actually co-refer (i.e to the dot-object lunch), but the predicates 'being delicious' and 'taking forever' apply to different facets of the object (i.e. respectively to the food facet and to the social event facet).

According to Recanati, fictional names are also polysemous (i.e. they can refer to flesh and blood individuals or to abstract objects) and denote dot-objects (e.g. the name 'Frodo' denotes the dot-object `flesh and blood individual • abstract object`). In metafictional statements the name 'Frodo' refers to the dot-object Frodo through its abstract object facet. In parafictional statements (both explicit and implicit) the name 'Frodo' refers to the dot-object Frodo through its flesh and blood individual facet.

Importantly, our concept[6] of the abstract object facet of Frodo "contains both nuclear information (the properties encoded by the fictional character) and extranuclear information (the properties exemplified by the fictional character)" [17, p.23]. What nuclear properties it contains is determined by our concept of the flesh and blood individual facet of Frodo (i.e. the concept of the abstract object facet contains a 'pointer' to the concept of the flesh and blood individual facet). In other words, Recanati includes Zalta's distinction between encoding and exemplifying properties (relevant for the abstract object facet of Frodo) and agrees that what properties are encoded is determined by our parafictional knowledge. But, whereas for Zalta Frodo the abstract object encodes just those properties that according to *The Lord of the Rings* Frodo the abstract object exemplifies; for Recanati, the abstract object facet of Frodo encodes just those properties that according to *The Lord of the Rings* the flesh and blood individual facet of Frodo exemplifies.

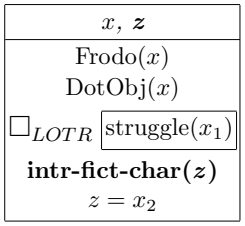
Applying Recanati's analysis to the workspace account suggests an adjustment of the fictive closure operation: At fictive closure we update the common ground with discourse referents for dot-objects for any newly introduced fictional character. These can be referred to as dot-objects (x), through their flesh and blood facet (x_1) or through their abstract object facet (x_2):

[6] In order to avoid metaphysical assumptions about the existence of multifaceted dot-objects, Recanati suggests that the correct objects of study are in fact dot-*concepts* (i.e. concepts of dot-objects) rather than dot-objects. In the DRS's below, a discourse referents and its associated conditions represent a concept (Cf. Maier [14])



Hence, x_1 as it appears in the parafictional statement, refers to the dot-object Frodo but through its flesh and blood individual facet.

A dot-object analysis of fictional characters solves the challenge posed by metafictional anaphora. Consider again our central example (5). I represent the common ground resulting from the parafictional update in (5) and the proposed metafictional update (in bold) as follows:



As the formalism shows the discourse referent x (predicated over through its abstract object facet in $z = x_2$) for 'Frodo' is accessible outside of the $LOTR$ fiction-operator. Hence we can equate the discourse referents for 'Frodo' and 'he' and interpret the metafictional statements in (5). Although the solution is formally very similar to the solution offered by an abstract account it avoids the problem identified with it because the name 'Frodo' in the explicit parafictional statement refers to the dot-object Frodo *through its flesh and blood facet* rather than to an abstract object.

Although a dot-object account of metafictional anaphora seems promising, some of the details still need to be worked out. First, as Recanati himself also notes ([17, p. 17, n.11]), the metaphysical status of dot-objects is unclear; Should we think of a dot-object simply as a pair consisting of two facets, i.e. $x = \langle x_1, x_2 \rangle$? Must all facets of a dot-object exist in order for the dot-object to exist? These questions are especially pressing in the case of fiction where one of the facets of the dot-object (i.e. the flesh and blood individual facet) does not exist. Second, it is not obvious how the crucial difference with an abstract object account (i.e. referring to a dot-object through different facets versus referring to an abstract object using different kinds of predicates), should be formalized. Whereas using the same argument (x) in the formalisation of metafictional and parafictional statements fails to show the difference, using different arguments (x_1 and x_2 as in the DRS above) suggests that fictional names in parafictional and metafictional statements refer to distinct objects rather than to one dot-object.

4 Conclusions

I have argued that a workspace account of fictional and parafictional statements runs into difficulties with metafictional anaphora because the discourse referent for the fictional name introduced in the parafictional statement is not accessible outside of the fiction-operator. I have evaluated three different accounts of metafictional anaphora: A descriptivist approach (that requires an additional account of how to accommodate appropriate definite descriptions), an abstract object account (that offers a solution to the problem of metafictional anaphora but wrongly analyses explicit parafictional statements as being about abstract objects) and a dot-object account (that solves the aforementioned problem but remains unclear on some crucial parts).

In this paper I have primarily focussed on pronominal anaphora across mixed parafictional/metafictional discourse. In fact, pronominal anaphora could occur across all possible types of mixed discourse with fictional, parafictional and metafictional statements (though some possibilities seem unlikely to actually appear such as pronominal anaphora across metafictional/fictional discourse). Eventually, an adequate account of fictional names will have to give the right predictions for anaphora across all acceptable types of mixed discourse (rather than only being able to account for statements in isolation).

More specifically, it would be interesting to extend the described accounts of metafictional anaphora to certain problematic cases. For instance, suppose that apart from *The Lord of the Rings* Tolkien also wrote an alternative story (*The Lord of the Schmings*) in which the character Gimli (a dwarf in *The Lord of the Rings*) is an elf. I could then felicitously say:

(12) In *The Lord of the Rings*, Gimli_i is a dwarf but in *The Lord of the Schmings*, he_i is an elf.

This is an example of pronominal anaphora across parafictional statements about different narratives. Intuitively, although the pronoun 'he' is anaphoric on the name 'Gimli', the terms do not refer to the same Gimli since he is ascribed inconsistent (individual-level) predicates in the two different narratives. This is reminiscent of both the phenomenon of counterfictional imagination (See e.g. Friend [6]) and Geach's Hob-Nob puzzle:

(13) Hob thinks a witch_i blighted Bob's mare, and Nob thinks she_i killed Cob's sow. (adapted from Geach [7])

Here the pronominal anaphora occur across two different propositional attitude reports and although the pronoun 'she' is anaphoric on 'a witch', there need not be one particular witch that is the object of thought of both Hob and Nob. Future research will have to determine how to account for (12) and determine its relation to other puzzles. A possible strategy, in Recanati's dot-object account, would be to claim that 'he' and 'Gimli' in (12) refer to a dot-object with three or four different facets: two for the flesh and blood facets for Gimli the dwarf and Gimli the elf and one or two abstract object facets (depending on whether we allow for inconsistent abstract objects).

5 Acknowledgements

This research is supported by the Netherlands Organisation for Scientific Research (NWO), Vidi Grant 276-80-004 (Emar Maier).

References

1. Asher, N.: Lexical meaning in context: A web of words. Cambridge University Press, Cambridge (2011)
2. Currie, G.: The nature of fiction. Cambridge University Press, Cambridge (1990)
3. Eckardt, R.: The semantics of free indirect discourse: How texts allow us to mind-read and eavesdrop. Brill Publishers, Leiden (2014)
4. Elbourne, P.: Situations and individuals. MIT Press, Cambridge, MA (2005)
5. Evans, G.: Pronouns, quantifiers and relative clauses (I). Canadian Journal of Philosophy **8**(3), pp. 467–536 (1977)
6. Friend, S.: The great beetle debate: a study in imagining with names. Philosophical Studies **153**(2), pp. 183–211 (2009)
7. Geach, P.: Intentional identity. Journal of Philosophy **64**, pp. 627–632 (1967)
8. Heim, I.: E-type pronouns and donkey anaphora. Linguistics and Philosophy **13**, pp. 137–177 (1990)
9. Kamp, H.: A theory of truth and semantic representation. In: Groenendijk, J.A.G., Janssen, T.M.V., Stokhof, M.B.J. (eds.) Formal methods in the study of language, Part 1., pp. 277–322. Blackwell Publishers Ltd, Oxford, UK (1981)
10. Klauk, T.: Zalta on encoding fictional properties. Journal of Literary Theory **8**(2), pp. 234–256 (2014)
11. Kripke, S.: Vacuous names and fictional entities. In: Philosophical troubles: Collected papers **1**, pp. 52–74. Oxford University Press, Oxford (2011)
12. Lewis, D.: Truth in fiction. American Philosophical Quarterly **15**(1), pp. 37–46 (1978)
13. Luo, Z.: Formal semantics in modern type theories with coercive subtyping. Linguistic Philosophy **36**(6), pp. 491–513 (2012)
14. Maier, E.: Attitudes and mental Files in Discourse Representation Theory. Review of Philosophy and Psychology **7**(2) pp. 473–490 (2016)
15. Matravers, D.: Fiction and narrative. Oxford University Press, Oxford (2014)
16. Pustejovsky, J.: The generative lexicon. MIT Press, Cambridge, U.S. (1995)
17. Recanati, F.: Fictional, metafictional, parafictional. In: Proceedings of the Aristotelian society **118**(1), pp. 25–54 (2018)
18. Semeijn, M.: A Stalnakerian analysis of metafictive statements. In: The proceedings of the 21st Amsterdam colloquium, pp. 415–424. (2017)
19. Stalnaker, R.C.: Pragmatics. Synthese **22**(1-2), pp. 72–289 (1970)
20. Stokke, A.: Lying and asserting. Journal of Philosophy **110**(1), pp. 33–60 (2013)
21. Walton, K.L.: Mimesis as make-believe: On the foundations of the representational arts. Harvard University Press, Harvard (1990)
22. Zalta, E.N.: Abstract objects: An introduction to axiomatix metaphysics. Springer, New York (1983)
23. Zalta, E.N.: Intensional logic and the metaphysics of intentionality. MIT Press, Cambridge, U.S. (1988)

Perspective blending in graphic media

Sofia Bimpikou

University of Groningen, Netherlands
s.bimpikou@rug.nl

Abstract. This paper discusses the representation of perceptual events in comics. I present "blended" pictures in which the experiencing character and her non-veridical perception are both represented from an external perspective in a single image. Inspired by Abusch & Rooth's (2017) analysis of free perception sequences and their modelling of veridical and non-veridical perception, I develop some proposals to model interpretation of non-veridical perception in blended pictures. I also discuss to what extent blended-perspective pictures are parallel to free indirect discourse in literature.

Keywords: picture semantics · pictorial narrative · perception · mental states · perspective shifting

1 Introduction

Is it possible for someone to have full access to another individual's perceptual experience? This may sound as fiction in real life, but in fiction, it is a common phenomenon. The pictures below, taken from Grant Morrison's comic book *Joe the Barbarian* and the cartoon *BoJack Horseman* illustrate this:

Fig. 1. (a) Image from graphic novel *Joe the Barbarian*, by Grant Morrison. (b) Snapshot from *Bojack Horseman* animated series, season 1 episode 11 *Downer Ending*.

In figure (1a), we see a character surrounded by his own hallucinations: here, Joe's toys have come to life. But the character is also depicted, which implicates that a narrator can 'see' both the character and his hallucinations from some external position. Similarly, in (1b), the protagonist, Bojack, depicted on the left

of the picture, is under the influence of drugs and sees himself in the mirror as a real horse. We can simultaneously see Bojack and the image of himself in the mirror as he perceives it. I will call such instances "blended-perspective" or simply "blended" pictures, as they allude to the narrator's geometrical perspective from which the scene is projected, but also to the character's internal reality or epistemic perspective.[1]

In this paper, I present some proposals to deal with blended pictures in visual narratives by using tools from semantic theories applied to language. In section 2, I give an overview of previous work on semantics of pictures; in section 3, I discuss three ways to analyse blended pictures; finally, in section 4, I conclude and point out some issues for further research.

2 Background

Pictures convey their own content. This has led to the development of semantic analyses of pictures in recent work (see Greenberg 2011, Abusch 2012, 2015, Abusch & Rooth 2017). These approaches extend the possible worlds semantic framework to pictures, based on the idea that, like sentences in language, pictures express propositions, i.e. sets of worlds. Greenberg (2011) suggests that the content of a picture is the set of scenes (i.e. worlds at a time and a location) the picture accurately depicts. Accordingly, a picture is an accurate depiction of a scene if it can be derived from the scene by specific rules of geometrical projection.

In general, geometric projection is defined as $\pi\{w, v, l, M\} = p$, meaning that a world w (at time t)[2] is projected to picture p from viewpoint v given a marking rule M and a line projection rule l.[3] The main idea is that the content of a picture p is a set of worlds relative to a geometrical viewpoint. In other words, the content of a picture p is a set of pairs consisting of a world and a viewpoint:

$$[\![p]\!]^{M,l} = \{\langle w, v\rangle | \pi(w, v, l, M) = p\} \tag{1}$$

Before moving to their analysis of looking events, I will say a few words about how discourse referents are identified in pictures. I follow Abusch's (2012) analysis of co-reference across panels. Abusch uses a Discourse Representation Theory framework to model co-reference of individuals in picture sequences. The reader

[1] Note that the term "blending" as used here is different from the notion of "blending" as used in cognitive semantics literature.

[2] The time parameter t is neglected in the formalisations, but "world" will refer to a "world at a time".

[3] M and l are parameters of geometrical projection. More specifically, l determines the projection lines from a viewpoint v towards a scene and M determines how points in the picture plane are to be marked with respect to the projection lines and the scene. For more details, see Greenberg (2011), Abusch (2015). Except for formula (1), I will not include these parameters in the following formalisations in order to keep things simple.

distinguishes certain areas in the picture that correspond to the story characters. Discourse referents are made out of these areas. Identity relations between discourse referents across pictures are then formalised as identity predications between these areas. Abusch suggests that co-reference is done at a post-semantic level, that is, identity between areas in pictures is determined pragmatically and is not part of the literal content of a picture.

Within a geometrical projection framework, Abusch & Rooth (2017) analyse free perception picture sequences. Their analysis is directly relevant for the data presented here. A free perception sequence (p, q) is a sequence in which one picture (p) depicts a character looking at a scene and the other one (q) - the free perception panel - depicts the scene looked at, as if directly through the character's eyes. Figure (2) is an illustration of a free perception sequence:

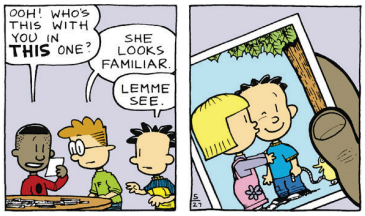


Fig. 2. A free perception sequence. (taken from *Big Nate*, by Lincoln Peirce)

A crucial distinction is that between veridical and non-veridical looking events. In the case of veridical looking events, the free perception panel has an extensional interpretation: it shows what the "base world" looks like from the character's geometric perspective, as in the picture above.[4] On the other hand, in the case of non-veridical perceptual events, the free perception panel depicts what the character sees but that may not correspond to how the base worlds looks like. This happens in cases of misperception, for example, when the protagonist hallucinates. This is shown in figure (3). Bart is looking at a jar which is actually empty, but what he sees instead is a dead fairy.

The authors propose covert syntactic embedding for free perception panels, inspired by natural language embedding structures.[5] More specifically, they pro-

[4] I assume that the "base world" corresponds to the objective representation of the fictional world by the reliable narrator. In pictorial narratives, as opposed to language, there is not always a verbal narrator. In the absence of a verbal narrator, I assume an impersonal narrator or a camera eye. How the notion of the narrator in visual media should be construed is a matter of debate in narratological studies. I will not take a stance on this debate here, but no matter how a narrator is construed, I take the default visual observer to be a reliable one. Therefore, the base world is the representation of the fictional world according to that reliable observer.

[5] They also suggest that extensional free perception panels can be analysed as top-level conjuncts but this option will not be discussed here.

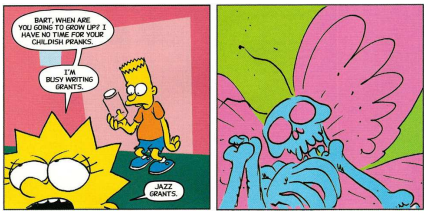


Fig. 3. A non-veridical free perception sequence. (taken from *Bart Simpson's Treehouse of Horror*, by Kyle Baker)

pose syntactic embedding of the free perception panel under a covert operator P, a *see*-predicate. This is illustrated in figure (4). The authors propose the following logical form for embedding cases (here simplified):

$$w, v, O \models p\, d\, [1[P\, q]] \tag{2}$$

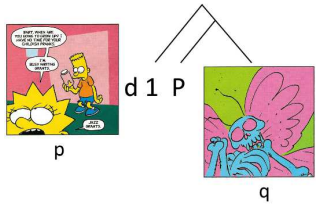


Fig. 4. An embedding structure for a free perception sequence (p, q).

Abusch & Rooth use a dynamic semantics framework. The above formula expresses a satisfaction clause where the tuple on the left of the turnstile satisfies the sequence (p, q) on the right. w is a world, viewpoint v is the viewpoint for the last picture of the sequence and O is a sequence of individuals onto which the discourse referents are mapped. On the right side of the turnstile, p and q are the two pictures and d is a discourse referent with index 1 introduced in picture p. According to the intensional LF (2), q is syntactically embedded under the covert operator P that takes the index introduced by the discourse referent as its subject. This means that w looks like q from agent $d1$'s perspective, but the base world may or may not look like q.

Note that this LF also allows for an extensional interpretation, as is the case in natural language embedding structures with verbs like *see* or *believe*. To capture ambiguity, the authors distinguish between veridical and non-veridical looking events: $l(x, q)$ and $m(x, q)$ respectively, that both translate into "x looks at a scene that projects to picture q from x's perspective", but the difference between them is that m has a precondition that the base world does not actually

look like q, only x sees it as q. Roughly, the idea is that an embedding structure entails that the agent has looked and that w looks like q from his perspective, but the base world might either look like q (therefore the world ends up with a veridical looking event l) or it may not look like q (hence the world ends up with a non-veridical looking event m).

Overall, the authors offer a neat proposal in order to allow for both extensional and intensional interpretations in free perception sequences. In the following section, I will use the main idea of their proposal to account for blended-perspective pictures.

3 Perspective blending: exploring solutions

3.1 First proposal: Splitting & viewpoint-shifting

Although Abusch & Rooth's (2017) analysis can account for free perception sequences, it is not clear how it could work for pictures like (1a): these seem like free perception sequences that are merged or blended into a single image. In this section I explore the idea of "unblending" such pictures by turning them into free perception sequences, and following a similar embedding as proposed in Abusch & Rooth (2017).

When seeing pictures like (1a), and of course based on the previous narrative, we infer that the scene surrounding the figure of the protagonist, call him j, reflects not the 'objective' world of the fiction, but the subjective world of the protagonist, that is, the world as perceived by the protagonist. In that case, I assume that the reader re-analyses the picture as something similar to a free perception sequence in the following way. First, a picture p is covertly split in two parts resulting in a sequence of two pictures: the first picture, call it p_1, contains the figure of the character, and the second picture, p_2, includes the whole scene. I postulate a splitting function f whose definition is given below and results in the sequence shown in figure (5).

Definition 1. *A splitting function f applied to picture p yields a sequence of two pictures p_1 and p_2: $f(p) = (p_1, p_2)$, where p_1 includes the figure of a salient discourse referent and p_2 includes the whole scene.*

How does this splitting take place, that is, how is each picture determined to contain what it contains? Since Joe is the salient protagonist in the preceding narrative, I assume that the covert splitting is the result of some pragmatic mechanism that can isolate a figure corresponding to a salient discourse referent and separate this from the rest of the picture (see the discussion on discourse referents in section 2). Joe is the salient protagonist of the story, the most prominent individual around whom the story revolves, thus, it is his epistemic state that is relevant for the interpretation of the non-veridical content of picture p.

The sequence created by the splitting function is not yet a free perception sequence: the second panel represents an external, third-person perspective, not an internal, first-person one. In order to use an intensional operator like Abusch

Fig. 5. Output of splitting function f: p_1(left), p_2 (right)

& Rooth's, we have to adjust the perspective of the second picture, p_2, because we want to capture the fact that the world in p_2 is the world as perceived by the character. In order to accommodate this, we can hypothesise that a second function operates which adjusts the current, external viewpoint to the character's viewpoint. The result is a set of hypothetical pictures with a first-person, internal perspective (i.e. a set of free perception panels), that 'match' the content of p_2.

The reason why the viewpoint-shifting function results in a *set* of pictures and not in a unique picture is because we do not really know what the character precisely sees: we just come to imagine what the world would look like from his point of view, therefore many pictures could be compatible with what he saw. Suppose picture (6a) below is the initial, external-perspective picture. Pictures (6b) and (6c) depict two possible worlds viewed from the character's eye location. Picture (6d) does not depict a possible world: in (6a), there is no big blue rabbit between the cube and the triangle. On the other hand, from picture (6a) we cannot see what the character sees precisely, therefore there could be nothing between the cube and the triangle (picture 6b) or there could be some small rabbit that is not visible from above (picture 6c). Hence, the viewpoint-shifting function gives us all and only the scenarios that match with picture (6a).

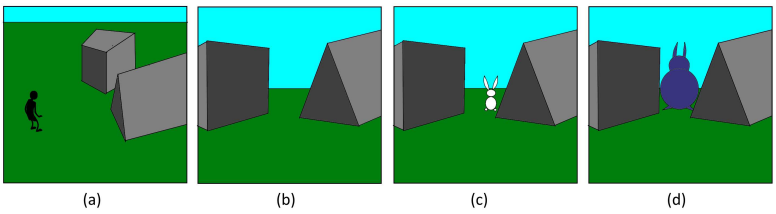


Fig. 6. (a) External (third-person) perspective picture. (b), (c), (d) Internal (first-person) perspective pictures.

The main idea is that a world can be represented from different viewpoints. In our case, this means that p_2 and the output pictures are different representations of the reality as perceived by the character. To make this precise, we abstract

away from viewpoints and use the definition of the uncentered content of a picture p (see Rooth & Abusch 2017):

$$[\![p]\!]^* = \{w \mid \exists v.\pi(w, v) = p\} \tag{3}$$

In words, the uncentered content of a picture p is the set of worlds w such that from some viewpoint v they are projected onto p.

Now we can give the definition of the viewpoint-shifting function:

$$g(v_j, p_2) = \{\pi(w, v_j) \mid w \in [\![p_2]\!]^*\} \tag{4}$$

Function g applies to the viewpoint of the discourse referent j introduced in p_1, i.e. v_j, and to picture p_2, and shifts the viewpoint of p_2 to v_j, yielding a set of first-person perspective pictures representing worlds that belong to the propositional, uncentered content of p_2. More simply, the output of g is a set of pictures $\{q_1, q_2...q_n\}$.

What we have now is actually a set of free perception sequences, whose first panel is p_1 and the second panel is a member of the output set of g, i.e. $\{q_1, q_2...q_n\}$. The remaining process is as proposed in Abusch & Rooth (2017), namely embedding each member of the set $\{q_1, q_2...q_n\}$ under an intensional operator P.

The proposal presented in this section is somewhat complex as it involves different steps. Is there a more simple analysis to model interpretation of blended pictures? In the following two sections, I consider two more ways.

3.2 Second proposal: Perspective blending as free indirect discourse?

A different approach would be to regard blended pictures as instances of free indirect discourse. In this section, I explore this idea.

Free indirect discourse is a literary style through which a character's thoughts and perceptual experience are represented with the mediation of the narrator's voice. This "intermediate mode" (McHale 2011) creates ambiguity as it is not clear if the point of view expressed is the narrator's or the character's. The usual pattern in free indirect discourse is that pronouns and tenses behave as in indirect discourse and reflect the narrator's perspective: third-person pronouns refer to the character(s) and past tenses refer to the character's present (at least in English). The rest of the expressions, such as temporal and locative adverbials (*now*, *here*), reflect the character's perspective (for a thorough overview of the expressions used in free indirect discourse, see Banfield 1982). For an illustration, see the example below:

(1) Tomorrow was Monday, Monday, the beginning of another schoolweek! (Lawrence, *Women in Love*, p. 185, as cited in Banfield 1982: 98)

In this example, *tomorrow* refers to the day following the day where the protagonist is temporally located, while the past tense *was* is anchored to the narrator's context.

Certain studies in the field of narratology discuss possible parallels of free indirect discourse in graphic novels and films (for example, Forceville 2002, Mikkonen 2008, Ghaffary & Nojoumian 2013). According to these studies, analogues of free indirect discourse in comics and/or in films are to be found in instances where the reader/viewer cannot determine whether what is represented visually (but also aurally in the case of films and animation) corresponds to the narrator's or the protagonist's perspective. For instance, free perception panels in comics and 'point-of-view' shots in film are taken in certain cases as possible parallels to free indirect discourse whenever they create ambiguity. But can this be supported semantically?

In semantic literature, different analyses are proposed about the status of free indirect discourse (see Banfield 1982, Schlenker 2004, Sharvit 2008, Eckardt 2014, Maier 2015, among others). Here I will remain agnostic as to which analysis is the most tenable. What is of special relevance is the behaviour of indexical expressions in free indirect discourse, which seems to be one of its defining characteristics. In standard discourse, the context of utterance is responsible for the interpretation of all indexical expressions such as first- and second-person pronouns, and also temporal and locative adverbials like *here* and *now*. Direct discourse can be considered a context-shifting mechanism because all expressions in a direct discourse report refer to the context of utterance being reported. By contrast, as mentioned above, in free indirect discourse, indexicals do not behave uniformly. This could lead to the assumption that free indirect discourse is basically a context-shifting mechanism for certain expressions.

Now, let's move to pictures. Here, there is no visual parallel of context-dependent linguistic expressions like temporal adverbials. However, change in the geometrical perspective of a picture is change in the locational point of reference, so this could be considered parallel to context-shifting. Free perception sequences involve change of perspective from one picture to another, so they can be viewed as instantiations of context-shifting. On the other hand, single pictures like (1a) do not involve any shifting as for a single picture there is by default only one corresponding viewpoint from which the whole scene (the character and the rest) is depicted (unless there are other embedded pictures).[6] Hence, pictures like (1a) do not seem to be semantic parallels to free indirect discourse if this split of indexicals is its defining feature.

However, in line with the observations in narratological studies, as far as the effect on the reader is concerned, blended pictures seem to convey the same ambiguity as passages in free indirect discourse do. More specifically, free indirect discourse reports are usually 'free' in that they are not embedded under an attitude or saying verb. Of course they can include a parenthetical verb indicating

[6] One general remark should be made. Throughout a graphic narrative, there is continuous switching between multiple viewpoints. However, I reduce my discussion to *two* viewpoints in order to refer to two broader notions: a) an external viewpoint that corresponds to the possibly multiple locations in space-time that the narrator can take, and b) an internal viewpoint that corresponds to a protagonist's first-person perspective.

whether the sentence is a speech or thought event (example 2 below), but only optionally. The absence of such a verb (example 3) may make the reader wonder if what is described is 'uttered' by the narrator and is therefore true in the story, or if the sentence only represents a character's thought or perceptual experience. See the examples below:

(2) It was seven o'clock, he thought.

(3) It was seven o'clock. (examples from Banfield 1982:205)

The same challenge is placed onto the reader of a graphic novel when seeing (1a): is this picture a representation of the actual world in the story or is the character hallucinating?[7]

To sum up, regarding the effect on the reader's interpretation, blended pictures have a similar impact as passages in free indirect discourse in novels and they too appear to be 'syntactically free'. Nonetheless, from a semantic point of view, such a view cannot be supported. In the following section, I will discuss a more plausible analysis.

3.3 Third proposal: Blended pictures as indirect discourse

In this section, I will explore an alternative analysis that makes use of an intensional belief-operator.

Instead of following the decompositional approach presented in section 3.1, we can suggest a simpler analysis. Pictures like (1a) can be regarded as parallel to indirect thought reports in language like "Joe thinks that he is surrounded by superheroes". In indirect discourse in written/spoken language, the narrator's perspective in the embedded clause is reflected through the use of the 3rd person pronoun. Something similar happens in pictures like (1a): the protagonist is also represented from a third-person perspective in the image. I assume that via inferential reasoning the reader will come to realise that what is going on in the picture is actually a hallucination. Inspired by Eckardt (2014), I assume a 'cautious update' for blended-perspective pictures (Maier & Bimpikou 2018).

The idea behind cautious update is that, even in normal, non-fictional discourse, we do not always update the common ground directly with the propositional content p of a sentence. If, for example, we consider the speaker confused, we take p to be part of the speaker's belief state only and not to form part of the shared common ground. Thus, instead of updating the common ground with the set of worlds where proposition p is true, we update with the proposition that the speaker believes that p.

Extending this to fictional discourse, a cautious update may take place whenever the reader assumes that a certain proposition is not true in the world of the fiction but true according to a protagonist, i.e. true in the protagonist's belief

[7] Ambiguity can be resolved through text, but here I just consider cases where no captions are included.

or imagination worlds. When seeing (1a), the reader infers that the protagonist hallucinates and so she has to perform a cautious update, i.e. embed the propositional content of the picture under an intensional belief-operator BEL. The update will result in interpreting the picture as something like "Joe believes that he is surrounded by superheroes". This results in the picture being interpreted as depicting the character's subjective world and not the actual world of the fiction. More generally, for a picture p and a salient protagonist j, $\text{BEL}_j p$ is true iff for all worlds $w\prime \in Bel_j$ (where Bel_j is j's belief state), $w\prime \in [\![p]\!]^*$ (here we use the uncentered, classical propositional content as defined in formula 3).

Should we appeal to a different operator, e.g. an imagination operator IMG? It should be remarked that imagination is different from hallucination or, more generally, faulty perception. When imagining, for example, when engaging in role playing or when daydreaming, we do know that our imaginary worlds are different from the actual, real world. On the other hand, in the case of faulty perception, there is no such awareness on the part of the perceiver. Therefore, imagining is distinct from misperceiving: the first involves the (aware) construction of a mental representation on the part of the agent, whereas misperception involves no distinction on her part between the actual and the imaginary world. Hence, when misperceiving, the agent actually believes that what she perceives is true. This is why a belief operator seems more appropriate. What is common in both imagining and believing though is that there are two different 'layers', the external and the internal reality. As for how the perceiver's awareness of the distinction between actual and mental is conveyed in each case through pictures, there seems to be a difference in marking in comics, as illustrated contrastively in figure (7). The blended picture (1a) is repeated in (7a); figure (7b) is a made-up image where the character appears instead with a thought bubble. Thought bubbles are conventionally used for imaginings and thoughts, so the most natural interpretation for figure (7b) would be that Joe is consciously thinking or imagining something. By contrast, for hallucinations there is no overt marking enclosing the character's perception and we can either have blended pictures (fig. 7a) or free perception panels (fig. 3).[8] So figures (7a) and (7b) seem to prompt different interpretations. Any particular choice (overt embedding with a bubble or non-embedding) has a significant effect on the reading process and consequently on the reader's interpretation. This makes the prediction that a picture with a thought bubble should be unambiguous, whereas pictures like (7a) can be ambiguous: a reader might fail to understand that what is represented in the picture is true only in the character's mind.

Although I mainly discuss single pictures, it is very common in comics to have sequences of blended pictures spanning a large part of the narrative, as in Bill Watterson's comic series *Calvin and Hobbes*. What would be the most satisfactory proposal from the ones suggested so far, also from the point of view of cognitive processing? According to the first proposal, the reader has to re-imagine the scene from a first-person perspective. The indirect discourse approach in-

[8] For a detailed discussion on speech and thought bubbles in comics and a somewhat different approach to their relation to "awareness", see Cohn (2013).

Fig. 7. (a) Blended picture. (b) Picture with a thought bubble.

volves the insertion of a belief operator without applying extra operations such as a viewpoint-shifting mechanism. For sequences of blended pictures, we may assume that these are grouped together as a constituent and that the intensional operator scopes over the whole constituent. This is also relevant for animation and film. Consider the movie *Fight Club* or the episode of *Bojack Horseman* (fig. 1b) where Bojack hallucinates. It is hard to imagine how splitting and reorienting suggested in 3.1 for pictures could be applied in continuous shots. For animation, we could suggest that the intensional operator could apply at the level of a whole scene (taking a scene to correspond to a series of successive shots that represent a certain spatio-temporal slice of the fictional world).

Overall, blended pictures can be paralleled to indirect discourse reports in language. That makes the indirect discourse analysis more appealing because it can apply to both pictorial and linguistic data.

4 Discussion

In this paper, I discussed perception representation in graphic narratives, mainly comics, and I focused on depictions of characters that are surrounded by their hallucinations. These data are similar to free perception sequences in that they also depict de se experience and therefore can also represent non-veridical perception. Unlike free perception sequences though, our data are single pictures. Our goal was to build on Abusch & Rooth's (2017) account in order to include these data as well. I proposed two ways to analyse blended pictures (sections 3.1 and 3.3) and suggested that a third option, namely comparing blended-perspective pictures to free indirect discourse (section 3.2), is not a tenable approach.

An interesting case is the representation of different kinds of perception in visual narratives. For instance, dreaming is a kind of perceptual experience that, on the one hand, is not exactly like thinking or conscious imagining and, on the other hand, it is not exactly like hallucination or misperception. The following questions arise: first, how are different kinds of perceptual experience conveyed in pictures and how are representations of perceptual events different or similar across different media? secondly, how are distinct perceptual phenomena in pictures to be modelled semantically? For instance, to the extent that hallucinating and misperceiving are different from dreaming, should we appeal to different

kinds of modal operators in the mental representations of blended-perspective pictures?

A different question is whether the above observations can be tested experimentally. I already suggested that blended pictures and pictures with bubbles represent different kinds of perceptual experience especially with respect to how they encode the agent's awareness of the imagined content. Do different ways of representation cause significant differences in readers' interpretations as was suggested in section 3.3? These issues are left for future work. I hope to have pointed out some interesting directions for further research.

Acknowledgements

I would like to thank my supervisor Emar Maier for his help and feedback, as well as the anonymous ESSLLI reviewers for their valuable comments. This research is part of Emar Maier's research project *The Language of Fiction and Imagination* supported by NWO Vidi Grant 276-80-004.

References

Abusch, D.: Applying discourse semantics and pragmatics to co-reference in picture sequences. Proceedings of Sinn und Bedeutung 17 (2012)

Abusch, D.: Possible worlds semantics for pictures. Handbook article. Draft (2015)

Abusch, D., Rooth, M.: The formal semantics of free perception in pictorial narratives. Proceedings of the 21st Amsterdam Colloquium (2017)

Banfield, A.: Unspeakable sentences. Routledge and Kegan Paul, Boston (1982)

Cohn, N.: Beyond speech balloons and thought bubbles: The integration of text and image. Semiotica **2013**(197), 35–63 (2013)

Eckardt, R.: The semantics of free indirect discourse: How texts allow us to mind-read and eavesdrop. Brill (2014)

Forceville, C.: The conspiracy in the comfort of strangers: Narration in the novel and the film. Language and Literature **11**(2), 119–135 (2002)

Ghaffary, M., Nojoumian, A.A.: A poetics of free indirect discourse in narrative film. Performance Studies: Rupkatha Journal on Interdisciplinary Studies in Humanities, Volume V, Number 2, 2013 p. 269 (2013)

Greenberg, G.J.: The semiotic spectrum. Rutgers The State University of New Jersey-New Brunswick (2011)

Maier, E.: Quotation and unquotation in free indirect discourse. Mind & Language **30**(3), 345–373 (2015)

Maier, E., Bimpikou, S.: Blending perspectives in pictorial narratives. Abstract accepted at Sinn & Bedeutung 23 (2018)

McHale, B.: Speech representation. `http://www.lhn.uni-hamburg.de/article/speech-representation` (2011), revised 2014 and accessed June 30, 2018

Mikkonen, K.: Presenting minds in graphic narratives. Partial Answers: Journal of Literature and the History of Ideas **6**(2), 301–321 (2008)

Rooth, M., Abusch, D.: Picture descriptions and centered content. Proceedings of Sinn und Bedeutung 21 (2017)

Schlenker, P.: Context of thought and context of utterance: A note on free indirect discourse and the historical present. Mind & Language **19**(3), 279–304 (2004)
Sharvit, Y.: The puzzle of free indirect discourse. Linguistics and Philosophy **31**(3), 353–395 (2008)

Free Relatives, Feature Recycling, and Reprojection in Minimalist Grammars

Richard Stockwell

University of California, Los Angeles
rstockwell15@ucla.edu

Abstract. This paper considers how to derive free relatives — e.g. *John eats [$_{DP}$ what Mary eats]* — in Minimalist Grammars. Free relatives are string-identical to indirect questions — e.g. *John wonders [$_{CP}$ what Mary eats]*. An analysis of free relatives as nominalised indirect questions is easy to implement, but empirical evidence points instead to wh-words reprojecting in free relatives. Implementing a reprojection analysis in Minimalist Grammars requires innovations to revise the stipulation that the probe always projects the head, and to allow features to be reused non-consecutively.

1 Introduction

This paper considers how to derive free relatives (FRs) (1) in Minimalist Grammars (MG) [15, 17]. Section 2 illustrates MGs with an analysis of indirect questions (IQs) (2), which are string-identical to FRs. An analysis of FRs as nominalised IQs is easy to implement, but the evidence presented in Sect. 3 points instead to wh-words reprojecting in FRs [7]. In order to implement a reprojection analysis of FRs, I propose two innovations to MG in Sect. 4: one, a Reproject operation that revises the stipulation that the probe always projects the head; and two, feature recycling, a way for features to be reused non-consecutively. I explore these innovations in Sects. 5 and 6 before concluding in Sect. 7.

(1) John eats [$_{DP}$ what he eats].

(2) John wonders [$_{CP}$ what Mary eats].

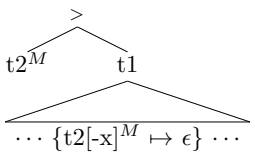
2 Minimalist Grammars, Indirect Questions, and Free Relatives

An MG analysis specifies a lexicon, pairing words with ordered lists of syntactic features. Matches between the first elements in these lists license applications of the structure building operations Merge and Move. We write *t[f]* when the head of a tree — found by following the headedness arrows $>$ and $<$ down to a leaf node — has a sequence of syntactic features whose first element is *f*, and *t* for

that tree with feature f erased. Merge (3) is licensed by matching category X and selector $=X$ features on the head of a pair of trees *t1* and *t2*. If the selector *t1* is lexical, it is linearized to the left $<$ and *t2* is called the complement; otherwise *t1* is linearized to the right $>$ and *t2* is called the specifier. Move (4) is licensed by matching probe $+x$ and goal $-x$ features on a tree *t1* containing a subtree *t2*. The probe *t1* takes as a specifier the maximal projection of *t2*, $t2^M$, which is made phonetically null in its original position.[1] The matching features that license Merge and Move are checked and deleted. Note especially that the selector/probe *t1* projects the head, whose remaining features drive further structure building.

(3) Merge (t1[=X], t2[X]) = $<$ (t1, t2) if t1 is lexical, $>$ (t2, t1) otherwise.

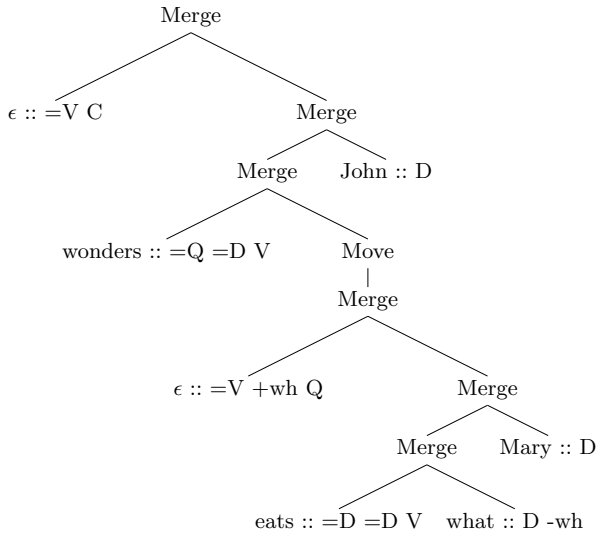
(4) Move (t1[+x]) =



An uncontroversial analysis of indirect questions (IQs) (2) can be implemented in MG with the lexicon in (5), illustrated with derivation (6) and derived (7) trees.

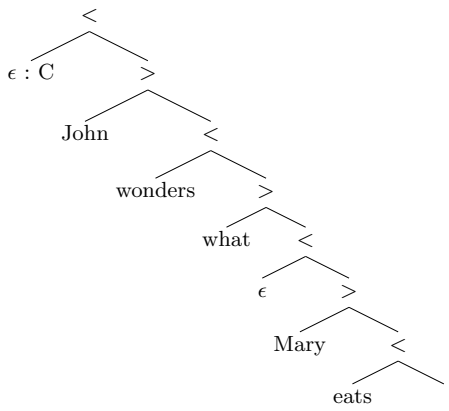
(5) John :: D Mary :: D wonders :: =Q =D V eats :: =D =D V
ϵ :: =V C ϵ :: =V +wh Q what :: D -wh

(6)



[1] Move is also subject to the shortest move constraint, which will not concern us here.

(7)



In two Merge steps, *eats* checks its $=D$ selector features against the category D features of *what*, then *Mary*. Omitting consideration of the tense layer, the null question complementiser $\epsilon :: =V\ +wh\ Q$ merges next. Move checks $+wh$ against $-wh$ to complete the embedded question. Construction of the main clause ends with the null complementiser $\epsilon :: =V\ C$ of the start category C.

The lexicon in (5) cannot derive free relatives (FRs) (1). Substituting *eats* for *wonder* does not converge, since *eats* selects for a complement of category D, not Q. A simple solution supplements the lexicon with $\epsilon :: =Q\ D$ — a null D that selects a Q complement. Merge of $\epsilon :: =Q\ D$ with the output of Move in (6) converts the IQ from category Q to D, which *eats* can then take as its complement. Several versions of this null head analysis of FRs have been proposed, e.g. [11, 12]. However, as shown in the next section, there is a good deal of evidence against it.

3 Dual Role of Wh-words in Free Relatives

On the null head analysis, the derivation of a FR proceeds via an IQ of category Q. Projecting Q and merging $\epsilon :: =Q\ D$ seals off the wh-word inside the IQ, preventing it from informing the rest of the derivation. However, the evidence suggests that the wh-word itself is the head of the FR, since the behaviour of a FR is keyed to the wh-word that forms it. This section shows that this is so for category distribution and case matching.

First, FRs distribute with the category of their wh-word (8), cf. [1]. A FR with *what* (1) distributes as a DP; but FRs formed with *where* distribute as PPs rather than DPs (8a), and those formed with *how* as AdvPs in not being able to intervene between a verb and its object (8b). On the null head analysis, this would require null, Q-complement-taking lexical items of many categories, e.g. $\epsilon :: =Q\ P$, $\epsilon :: =Q\ Adv$. And even then, nothing would enforce category matching between the null lexical item and the wh-word inside the FR, as required to rule out (9); in other words, we would expect mixtures like $what :: D\ -wh$ and $\epsilon :: =Q\ P$ to be grammatical.

(8) a. i. Mary put the book [$_{PP}$ on the shelf] / [$_{PP}$ where she keeps it].
ii. *Mary put the book [$_{DP}$ the shelf] / [$_{DP}$ what she built].
b. i. John speaks [$_{AdvP}$ quickly] / [$_{AdvP}$ how you speak].
ii. *John takes [$_{AdvP}$ quickly] / [$_{AdvP}$ how you write letters] notes.

(9) *Mary put the book [$_{DP}$ what John built].

Second, in languages with morphological case, e.g. German [13], the wh-word in a FR must satisfy the case requirements of both the relative and matrix clauses. (10) is grammatical, since the nominative wh-word is the subject of the FR, which is the subject of the sentence. But (11) is ungrammatical due to the competing case requirements placed on the wh-word inside the FR, where it is an accusative object, and the FR as a whole, which is the nominative subject of the sentence. This conflict cannot be resolved — neither the accusative nor the nominative form of the wh-word will do. Since the null head analysis involves two distinct lexical items of category D — *what* and ϵ :: $=Q\ D$ — it offers no explanation for why they should match in case.

(10) [$_{DP_{NOM}}$ Wer$_{NOM}$ nicht stark ist] muss klug sein.
who not strong is must clever be.
'Who is not strong must be clever.'

(11) *[$_{DP_{NOM}}$ { $^{Wen_{ACC}}_{Wer_{NOM}}$ } Gott schwach geschaffen hat] muss klug sein.
who God weak created has must clever be.
'Who God has created weak must be clever.'

Matches between FRs and their wh-words in distribution and case argue that the moving wh-word itself projects the head of FRs to determine their behaviour in the rest of the derivation. A number of researchers have reached this conclusion [1, 3, 6, 7]. However, the reprojection analysis directly contradicts the standard stipulation that a probe always projects over the goal it attracts to its specifier [5]. This stipulation is built into the definition of Move in (4), meaning MG cannot accommodate a reprojection analysis of FRs without amendments. The next section proposes a way to overcome this stipulation and implement a reprojection analysis of FRs in MG.

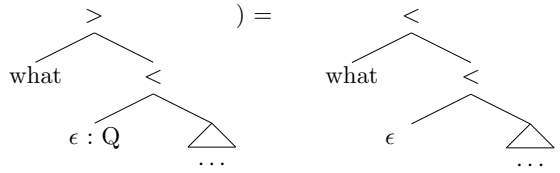
4 Implementing Reprojection in Minimalist Grammars

This section seeks to implement a reprojection analysis of FRs in MG. It does so by proposing two innovations: (i) Reproject, a new structure-building operation that revises the stipulation that the probe always projects; and (ii) feature recycling, a way for the category feature of the wh-word to be reused in the face of the resource sensitivity of MG. Consequences of these innovations and further directions are explored in Sects. 5 and 6.

4.1 Reproject

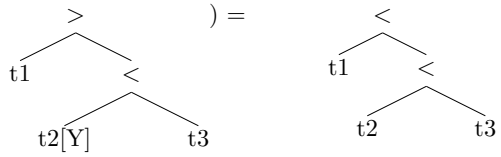
In revising the stipulation that a probe always projects, I propose to add Reproject to MG's inventory of structure building operations. We want Reproject to apply as in (12), reversing headedness to make *what* the head, thereby allowing *what* to determine the future of the derivation; and deleting the category feature Q, which would otherwise be left unchecked and cause the derivation to crash.[2]

(12) Reproject ( ) =



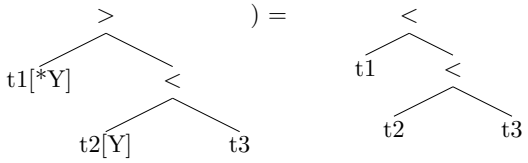
A general definition of Reproject is given in (13). A unary operation applying to a tree with specifier *t1*, head *t2*, and complement *t3*, Reproject switches headedness to *t1* and deletes the category feature of *t2*, leaving *t3* unchanged.

(13) Reproject ( ) =



However, the way Y is checked without matching another feature in (13) would make Reproject very different from Merge and Move, which symmetrically check pairs of matching features.[3] Reproject is defined symmetrically in (14), where it applies to a tree where a reprojection feature $*Y$ on the specifier *t1* matches the category of the head *t2*. Both features are checked, and headedness switches to *t1*. Using (14) means adding reprojection features $*Y$ to the inventory of syntactic features, and FR-specific reprojecting versions of wh-words to the lexicon; e.g. *what :: D -wh *Q*, *where :: P -wh *Q*.[4]

(14) Reproject ( ) =



[2] The question of what features are on *what* in (12) is postponed to the next subsection.

[3] Even with persistent features [16], as discussed in the next subsection, while checking is not necessarily symmetric, structure building is still licensed by pairs of matching features.

[4] Wh-clustering [10] — see Sect. 5.3 below — provides a precedent for Reproject in being triggered by a feature on a specifier rather than a head. Clustering also involves complex specifiers, whereas I restrict attention here to trees with exactly one specifier.

However, as things stand the outcome of (14) has no features. *t1* is the head, but all its features have been checked en route to it becoming the specifier of *t2*. In deriving the FR in (1), *what :: D -wh *Q* has its *D* checked by Merge with *eats*, *-wh* by Move, and **Q* by Reproject, rendering its feature list empty, i.e. *what* : ϵ. The next subsection proposes a way for category features to be reused so that the wh-word can serve as the head of FRs after Reproject.

4.2 Feature Recycling

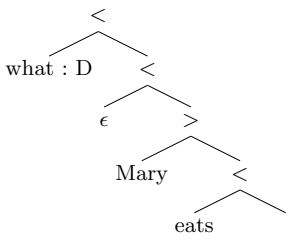
In order to account for the matching effects observed in Sect. 3, we would like *what* to play a dual role in deriving (1) by contributing its category feature twice: first as complement to *eats*; then again after Move and Reproject to categorise the FR. However, MG structure building is resource sensitive: the matching features that license Merge and Move are checked and deleted. *D* of *what :: D -wh *Q* is expended in Merge with *eats*, and is subsequently unavailable.

Endowing *what* with a second *D* feature ordered after *-wh* and **Q*, i.e. *what :: D -wh *Q D*, (cf. *where :: P -wh *Q P*) invites the same empirical challenges as the null head analysis: with two separate category features, nothing enforces category and case matching between FRs and their wh-words. Instead, we would like one and the same *D* feature to contribute twice to the derivation.

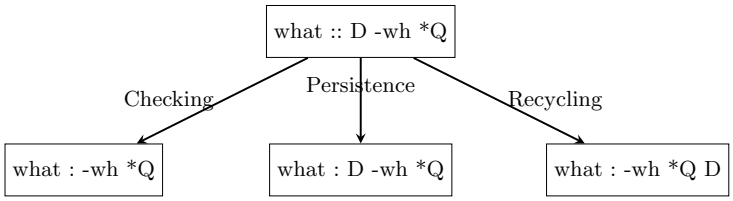
Persistent features are an existing innovation that allow category features to be used multiple times [16]. Merge continues to be licensed symmetrically by matching features, but persistent features (underlined $\underline{F}$) do not have to delete. Persistent features were motivated for implementing the movement theory of control [8], allowing the same $\underline{D}$ to occupy multiple argument positions by satisfying multiple *=D* features. However, persistence in *what ::* $\underline{D}$ *-wh *Q* does not help in deriving FRs, since the two desired uses of *D* are non-consecutive. Move and Reproject must apply after Merge of *what* with *eats* but before *what* categorises the FR. Hence $\underline{D}$ would have to delete to allow Move to be triggered by *-wh* before having the chance to provide the category of the FR.

I therefore propose feature recycling. Beyond persisting at the head of a list, features can live on in the derivation by cycling to the end of the list after licensing Merge. Feature recycling allows the *D* of *what :: D -wh *Q* to play a dual but non-consecutive role, as shown in the derived tree of the FR from (1) in (15). After licensing Merge with *eats*, *D* cycles to the end of the feature list; and after Move checks *-wh* and Reproject checks **Q*, the recycled *D* is back at the head of *what*'s feature list to serve as the head of the FR. The diagram in (16) summarises the differences between standard resource sensitive feature checking, persistent features, and feature recycling with respect to *what*'s *D* feature.

(15)



(16)



Thus we have implemented a reprojection analysis of FRs, cf. [1, 3, 6, 7]. However, the analysis has come at the cost of two innovations. The first, Reproject, reverses headedness to the wh-word and deletes the category feature of the embedded clause, which would otherwise be left unchecked and cause a crash. The output of Reproject would lack any features were it not for the second innovation, feature recycling, which provides a way for the wh-word's category feature to be reused non-consecutively as the head of the FR. I explore these innovations further in the next two sections.

5 More on Reproject

This section considers the Reproject operation in greater detail, with discussion of the location of the triggering feature, Reproject's relationship to Move, and multiple-wh FRs.

5.1 Wh-word Trigger

This subsection attempts to justify making the wh-word the trigger for Reproject. Whereas we could have put the triggering feature **Y* on the head that is reprojected over, the definition of Reproject in (14) has the triggering feature on the wh-word, e.g. *what :: D -wh *Q*. The argument is based on the complement restriction on FRs. Wh-words can only form FRs if they lack a complement, as shown by the ungrammaticlity of (17).

(17) *John eats [$_{DP}$ what food Mary eats].

This restriction is much easier to state if the wh-word is the trigger for Reproject. We can exclude from the lexicon wh-words with both a selector feature

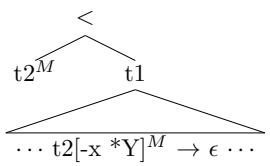
and a Reproject feature, e.g. **what :: =N D -wh *Q*. By the time ϵ *:: +wh Q* interacts with the wh-word in the Move step, on the other hand, it is unable to discriminate between a wh-word with or without a complement; in either case the wh-word will now have *-wh* as its first feature, *=N* having long since been checked. Since the relevant information to distinguish between good FRs and (17) is not available to ϵ *:: +wh Q*, the complement restriction on FRs argues that the wh-word should trigger Reproject.

However, this conclusion is provisional. Much stronger evidence would be cases of wh-words reprojecting over lexical items other than ϵ *:: +wh Q*. Only then could we be sure that the wh-word is the trigger for Reproject, rather than the head reprojected over.

5.2 Reprojecting Move

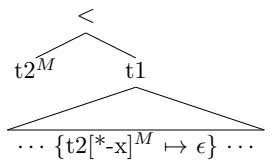
Reproject as in (14) involves the heads *what :: D -wh *Q* and ϵ *:: =V +wh Q*. It is licensed symmetrically by matching between the reprojection feature **Q* and category feature *Q*. But the two heads match in more than *Q* — they also match for *wh*. In Sect. 4.1, *wh* was checked by an application of Move before Reproject. But nothing said there enforces the co-occurrence in a lexical item's feature list of a Move licensee like *-wh* and a Reproject trigger like **Q*. If Reproject is always fed by Move, we would be missing a generalization. Instead, we could recast Reproject as a version of Move, as in (18).

(18) Reprojecting Move (t1[+x Y]) =



The definition in (18) enforces a dependency between Move licensees and Reproject triggers. Continuing to assume that the wh-word is the trigger for Reproject, we could further strengthen the dependency between moving and reprojecting by collapsing Move licensing and Reproject triggering into a single feature, **-x*. The reprojecting lexical item *what :: D *-wh* would then trigger Reprojecting Move, as defined in (19).

(19) Reprojecting Move (t1[+x Y]) =



However, the definition in (19), reintroduces the asymmetry problem from our first definition of Reproject in (13). The category feature *Y* on *t1* is asymmetrically checked, deleting without having matched with another feature. Beyond

this technical point, I cannot see how to decide between (18) and (19) as the definition of Reprojecting Move.

More generally, Reprojecting Move raises problems regardless of which of (18) or (19) we choose. For one, it increases the size of the 'moving window' on feature lists from one to two. Whereas Merge (3) and Move (4) apply based only on the first feature of the head, Reprojecting Move requires sight of the first two. A second problem might be that we have to restate all the restrictions on Move for Reprojecting Move, like the shortest move constraint and islands. Thus while reducing redundancy in enforcing a dependency between Move licensees and Reproject triggers, Reprojecting Move might increase redundancy elsewhere in the system. However, redundancy only arises to the extent that movement as it feeds reprojection is subject to the same constraints as ordinary movement. If empirical investigation finds it to be subject to different constraints, we would have a strong argument for Reprojecting Move as its own operation. As with the previous subsection, we end by wondering whether wh-words reproject over lexical items beyond $\epsilon :: +wh\ Q$.

5.3 Clustering and Reproject

Languages with overt multiple-wh movement to Spec,CP — like Bulgarian and Romanian (20) [2] — also allow multiple-wh FRs [14]:

(20) Ti-am dat [ce unde când a trebuit instalat].
CL2-have.1SG given what where when has needed installed
'I have given you the things that needed to be installed in the appropriate place at the appropriate time.'

Multiple-wh FRs distribute with the category of the topmost wh-word — *D* in (20). In terms of an MG analysis of wh-clustering [10], this means there is a dependency between Reproject triggers and Move licensees, but not Cluster licensees. Working with the definition of Reproject in (14), the lexical entries for the wh-words in (20) would be *ce* :: $D\ \triangledown wh$ *-wh* $^*Q$, *unde* :: $P\ \triangledown wh\ \triangle wh$ and *când* :: $P\ \triangle wh$. The reprojection trigger $^*Q$ co-occurs with *-wh*, not the Cluster licensor $\triangledown wh$ or licensee $\triangle wh$.

6 More on Reusing Features

Section 3 emphasised matches between FRs and their wh-words match in order to argue that the moving wh-word projects the head of FRs. Section 4.2 proposed feature recycling as a way for the wh-word's category feature to be reused to implement a reprojection analysis of FRs. This section explores ways in which a FR and the wh-word that forms it can behave differently — if only very slightly. Slight differences regarding case syncretism, complement/adjunct *where* FRs, and A-bar features suggest that the notion we need may not be recycling but refreshing, returning to the lexicon to pick another list of features compatible with the morphological form of the word.

6.1 Case Syncretism

Despite the disucssion of case matching in Sect. 3, I have yet to mention case features. In MG, all lexical items of category D also bear a k(ase) feature *-k*. The ungrammaticality of (21) shows that *-k* must recycle along with D in deriving FRs, since the FR as a whole must be in a case position.

(21) *It seems [$_{DP}$ what John eats] to be nice.

English wh-words do not differ morphologically for case,[5] which might suggest a generic *-k* feature for English rather than more articulated *-nom*, *-acc*, etc. Support for generic *-k* comes from the lack of case matching effects in English FRs. (22) is grammatical, despite the wh-word being assigned accusative internal to the FR, while the FR is nominative in the sentence overall. This contrasts with the German mismatch from Sect. 3, repeated here as (23).

(22) [$_{DP_{NOM}}$ What$_{ACC}$ John ate] killed him.

(23) *[$_{DP_{NOM}}$ { $^{Wen_{ACC}}_{Wer_{NOM}}$ } Gott schwach geschaffen hat] muss klug sein.
who God weak created has must clever be.
'Who God has created weak must be clever.'

(24) [$_{DP_{NOM}}$ Wer$_{NOM}$ nicht stark ist] muss klug sein.
who not strong is must clever be.
'Who is not strong must be clever.'

For English, then, we can say that generic *-k* of *what :: D -k -wh *Q* is licensed inside the FR by a case-assigner, recycled along with D, and licensed again by another case-assigner in the main clause. In German, on the other hand, the morphological differences among wh-words for case might suggest lexical items of category D differ among *-nom*, *-acc*, etc. In (24), *wer :: D -nom -wh *Q* is licensed for nominative inside the FR, recycled with D in forming the FR, and licensed again for nominative in the main clause.

Switching between *-acc* and *-nom* is ungrammatical in (23), but this is not always the case in German for a FR configuration that mixes nominative and accusative. Such mismatches are possible when there is case syncretism, as in (25) [13], where neuter gender *was* can realise either nominative or accusative.

(25) [$_{DP_{NOM}}$ Was$_{ACC}$ du gekocht hast] ist schimmlig.
What you cooked have is moldy.
'What you have cooked is moldy.'

Starting with *was :: D -acc -wh *Q* inside the FR, we cannot switch to *was :: D -nom -wh *Q* in the main clause via feature recycling, incorrectly predicting (25) to be bad. We could salvage feature recycling by changing our assumptions about case features, claiming that there is only one lexical item *was :: D -nomacc -wh *Q* whose underspecified *-nomacc* case feature can be

[5] I set aside *whom* as an archaism.

checked by either a nominative or accusative case-assigner. Alternatively, we could account for (25) by refreshing rather than recycling the features of *was*. After moving to specifier position and reprojecting, *was :: D -acc -wh *Q* has exhasted its list of features. Rather than pre-empting this problem with feature recycling, *was* : ϵ could refresh its features by reaching back into the lexicon for a list of features compatible with its morphological form. This refreshed list could be slightly different — including *-nom* rather than *-acc* in deriving (25).

While either underspecified *-nomacc* or feature refreshing would account equally well for German case syncretism, refreshing appears to be the only plausible option for the topic of the next subsection.

6.2 Complement vs. Adjunct *Where* Free Relatives

The previous subsection showed that syncretism allows wh-words and the FRs they form to differ in case. This section argues that sycretism also allows differences in category.

Following a prominent analysis of adjunction in MGs [9], PPs have very different categories depending on whether they appear in complement or adjunct position: *where :: P -wh* is a complement to verbs like *put :: =D =D =P*, whereas *where* :: $\approx V$ *-wh* adjoins to category *V*, which continues to be the head. In (26), *where* is an adjunct to *eats* inside the FR, while the FR as a whole is a complement to *put*. Thus *where* has different category features internal and external to the FR, which would not follow from feature recycling.

(26) Mary put the book [$_{PP_{comp}}$ where$_{PP_{adj}}$ John eats].

It is difficult to countenance an underspecification analysis among two different feature lists for *where* along the lines of underspecified *-nomacc* in the previous subsection. That leaves us with feature refreshing: in deriving (26), *where* :: $\approx V$ *-wh* is exhausted to *where* : ϵ in deriving the FR, before refreshing as *where :: P -wh* for the main clause.

6.3 A-bar features

Whether features are recycled or refreshed, A-bar features are not reused. Despite being headed by a wh-word, a FR cannot itself undergo wh-movement, as in (27).

(27) *[$_{DP}$ What John eats] does Mary eat *t*?

Unlike category and case features, which play a dual role in deriving FRs, *-wh* is definitively consumed in moving *what* inside the FR, so would have to be barred from recycling. In terms of refreshing, meanwhile, we could say that features are refreshed based on the non-wh part of the word, assuming decomposition of e.g. German *wer* into wh *w-* + *-er* nominative *D*.

Yet FRs can embark on other A-bar movements, e.g. topicalisation in (28).

(28) [$_{DP}$ What John eats], I eat *t*.

Still, the movement in (28) cannot result from reusing a feature. Assuming topicalisation is licensed by *-top*, it must be added to the FR after it has been fully formed: while the FR as a whole is topicalised in (28), the wh-word does not undergo topicalisation inside the FR, so *-top* cannot have been present on *what* at the start of the derivation. The opposite behaviour of *-wh* and *-top* in being active only internal vs. external to the FR tracks the difference between intrinsic vs. optional features [4, p. 231].

7 Conclusion

This paper set out to derive FRs in MG. Reviewing category and case matching effects motivated implementing a reprojection analysis. Doing so came at the cost of two innovations. Reproject, a new structure-building operation, revised the stipulation that the probe always projects. The technical questions of whether the trigger is the wh-word, and whether Reproject is a special case of Move, rest on the empirical question of whether wh-words reproject over lexical items other than $\epsilon :: +wh\ Q$. Feature recycling provided a way for the category feature of the wh-word to be reused nonconsecutively as the head of the FR in the face of the resource sensitivity of MG. The slight relaxation of matching effects where there is syncretism suggested features might be refreshed rather than recycled, though A-bar features cannot be reused.

References

1. Bresnan & Grimshaw (1978). The syntax of free relatives in English. Linguistic Inquiry 9(3): 331–91.
2. Caponigro & Fălăuş (2018). The functional nature of Multiple Wh- Free Relative Clauses in Romanian. Poster presented at SALT 28, MIT.
3. Cecchetto & Donati (2015). (Re)labeling. MIT Press.
4. Chomsky (1995). The Minimalist Program. MIT Press.
5. Chomsky (2008). On phases. In Foundational issues in linguistic theory: Essays in honor of Jean-Roger Vergnaud, 133–166.
6. Citko (2008). Missing labels. Lingua 118: 907–44.
7. Donati (2006). On wh-head movement. In Wh-movement: Moving on, 21–46.
8. Hornstein (1999). Movement and control. Linguistic Inquiry 30: 69–96.
9. Frey & Gärtner (2002). On the Treatment of Scrambling and Adjunction in Minimalist Grammars. Proceedings of Formal Grammar 2002.
10. Gärtner & Michaelis (2010). On the Treatment of Multiple-Wh-Interrogatives in Minimalist Grammars. Language and Logos, 339–366.
11. Groos & van Riemsdijk (1981). Matching effects in free relatives: A parameter of the core grammar. In Theory of Markedness in Generative Grammar, 171–216.
12. Grosu (1994). Three Studies in Locality and Case. Routledge.
13. van Riemsdijk (2007). Free Relatives. In The Blackwell Companion to Syntax .
14. Rudin 2007. Multiple wh-relatives in Slavic. FASL, 282–307.
15. Stabler, Ed (1997). Derivational minimalism. LNCS 1328: 68–95.
16. Stabler (2006). Sidewards without copying. Formal Grammar 11, 133–146.
17. Stabler (2011). Computational perspectives on minimalism. Oxford Handbook of Linguistic Minimalism, 617–642.