Proceedings of the
ESSLLI 2017
Student Session

29th European Summer School in Logic, Language & Information
July 17-28, 2017, Toulouse, France

Karoliina Lohiniva
Johannes Wahle
(Editors)
Preface

These proceedings contain the papers presented at the Student Session of the 29th European Summer School in Logic, Language and Information (ESSLLI 2017), taking place at the Toulouse Capitole University from July 17th to 28th, 2017. The Student Session is part of the ESSLLI tradition and was organized for the 22nd 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 Student Session attracted submissions from 19 different countries this year from all over Europe and beyond. As in previous years, the submissions were of high quality, and acceptance decisions were hard to make. We received 50 submissions, 35 of which were submitted for oral presentations, and 15 of which were submitted for poster presentations. At the Student Session, 14 of these submissions 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.

We would like to thank each of the co-chairs for all their invaluable help in the reviewing process and organization of the Student Session. Without them, the Student Session would not have been able to take place. Additionally, we would like to thank the area experts for their help in the reviewing process and their support of the co-chairs. We would also like to thank the ESSLLI Organizing Committee, for organizing the entire summer school, and catering to all our needs, and the program committee chair Shravan Vasishth for his support. Thanks go to the chairs of the previous Student Sessions, in particular to Ronald de Haan, Philip Schulz, Miriam Kaeshammer, Marisa Köllner and Ramon Ziai for providing us with many of the materials from the previous years and for their advice. As in previous years, Springer-Verlag has generously offered prizes for the Best Paper and Best Poster Award, and for this we are very grateful. 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.

July 2017

Karoliina Lohiniva & Johannes Wahle
Chairs of the ESSLLI 2017 Student Session
Organization

Program Committee

Chairs
- Karoliina Lohiniva (University of Geneva)
- Johannes Wahle (Universität Tübingen)

Language & Computation co-chairs
- Emmanuele Chersoni (Aix-Marseille University)
- Jennifer Sikos (Universität Heidelberg)

Logic & Computation co-chairs
- Herbert Lange (University of Gothenburg)
- Marie Farell (National University of Ireland Maynooth)

Logic & Language co-chairs
- Milica Denic (Ecole Normale Supérieure de Paris)
- Mora Maldonado (Ecole Normale Supérieure de Paris)

Area Experts

Language & Computation
- Nicolas Asher (IRIT Toulouse - Université Paul Sabatier)

Logic & Computation
- Samson Abramsky (University of Oxford)
- Jakub Szymanik (University of Amsterdam)

Logic & Language
- Sam Alxatib (City University of New York)
- Philippe Schlenker (Ecole Normale Supérieure de Paris)
# Table of Contents

## Logic & Computation

On One Variable Fragment of First Order Logic with Modulo Counting Quantifiers

*Bartosz Bednarczyk*

Towards Symbolic Factual Change in DEL

*Malvin Gattinger*

Certified Soundness of Simplest Known Formulation of First-Order Logic

*John Bruntse Larsen*

Lexicographical Update for Comparative Relation

*Kai Li*

Dialogue Games for Minimal Logic

*Alexandra Pavlova*

Resolution Calculi for Modal Logic and their Relative Proof Complexity

*Sarah Sigley*

## Logic & Language

Might Counterfactual Donkey Sentences

*Sam Carter & Simon Goldstein*

Towards a copular approach to Chinese clefts--Evidence from diachronic syntax

*Jun Chen*

The Pragmatics of Speaker Repeat Questions

*Gisela Desselkamp*

An additive ambiguity: summing events or degrees

*Cara Feldscher*
Restructuring and Actuality Entailment in Mandarin Obligatory Control

Yuyin He

Optional negative concord with Turkish neither..nor

Paloma Jeretič

An Analysis of Counteridenticals in Terms of Dream Reports

Carina Kauf

Uniformity Motivated

Cameron Domenico Kirk-Giannini

Deriving even though from even

Gunnar Lund

Where Force Matters: Embedding Epistemic Modals Under Doxastic Attitudes

Maša Močnik

Are Expressives Presuppositional? - The Case of Slurs

Tristan Thommen

Language & Computation

Exploring Compositionality of Estonian Particle Verbs

Eleri Aedmaa

The Effect of Poor Coreference Resolution on Document Understanding

Maximilian Droog-Hayes

Hierarchical Clustering of Estonian Verb Constructions

Dage Särg

In-domain or out-domain word embeddings? A study for Legal Cases

Milagro Teruel and Christian Cardellino
On One Variable Fragment of First Order Logic with Modulo Counting Quantifier

Bartosz Bednarczyk *

bbednarczyk@stud.cs.uni.wroc.pl
Institute of Computer Science
University of Wroclaw
Wroclaw, Poland

Abstract. We consider the one-variable fragment of first-order logic extended with modulo counting quantifiers. We prove \textit{NPTime}-completeness of such fragment by presenting an optimal algorithm for solving its finite satisfiability problem.

Keywords: finite satisfiability, computational complexity, decidability, classical decision problem, modulo counting quantifiers

1 Introduction

It is well known that first-order logic is not expressive enough to describe many natural properties including quantitative properties like parity. The last problem can be solved by adding to the language modulo counting quantifiers of the form "there exists a mod b elements x such that...". Such quantifiers were introduced by Wolfgang Thomas and coauthors in the 80s of the last century. A survey of main results in this area can be found in [16].

Recent research trends in first order logic extended with modulo counting quantifiers involve a study of definability of regular languages on words and its connections to algebra [16], [17]; definable languages of $(\mathbb{N}, +)$ [13], [8]; equivalences of finite structures [9]; definable tree languages [11], [3]; locality [7]; and extensions of Linear Temporal Logic e.g. [1], [14].

It is natural to ask whether decidable fragments extended with such quantifiers remain decidable and what is the complexity of their satisfiability problem. To our surprise, it seems that almost nothing is known in this subject. The only exception is the investigation of the complexity of two-variable fragment of first-order logic on finite words, carried out in the PHD thesis of A. V. Sreejith [15].

With this work we start a project aiming for a classification of decidable fragments of first-order logic extended with modulo counting quantifiers with respect to the complexity of their satisfiability problem. Natural candidates are the fragments with limited number of variables. Obviously if we allow to use at least three variables and such quantifiers, the satisfiability problem is undecidable, because it is undecidable already without modulo quantifiers. But the case of one or two variables is still open.

* Supported by the Polish National Science Centre grant No. 2016/21/B/ST6/01444.
In this paper we report on the one-variable case. More precisely, we prove \textit{NPTime}-completeness of the one-variable fragment of the first-order logic extended with modulo counting quantifiers by presenting an optimal algorithm for solving the finite satisfiability problem.

In the last section of this paper we briefly describe our ongoing research on some related fragments.

2 Preliminaries

We employ standard terminology from model theory and linear algebra.

2.1 Logic and Types

We refer to structures with fraktur letters, and to their universes with the corresponding Roman letters. We always assume that structures have non-empty universes. By $\mathbb{Z}_n$ we denote the set of all reminders modulo $n$, that is the set $\{0,1,\ldots,n-1\}$.

We define the one-variable fragment of first-order logic with modulo counting quantifiers $\text{FO}^1_{\text{MOD}}$ as the set of all first-order formulas (without functions, constants and relations of arity greater than one), featuring only the variable $x$, extended with modulo counting quantifiers of the form $\exists^\leq a (\bmod b)$, $\exists^= a (\bmod b)$ and $\exists^\geq a (\bmod b)$, where $b \in \mathbb{Z}_+$ is a positive integer and $a \in \mathbb{Z}_b$. The formal semantics of such quantifiers is as follows:

$$ M \models (\exists^\leq r (\bmod b) \varphi(x)) \overset{\text{def}}{\iff} \exists r \in \mathbb{Z}_b | \{x \in M : \varphi(x)\} | \equiv r (\bmod b) \land r \geq a, $$

where $\geq \in \{\leq, =, \geq\}$. We assume that numbers in quantifiers are written in binary system. Additionally, we allow $b$ to be the infinity symbol and we identify $\exists^\leq a (\bmod \infty)$ with the counting quantifier $\exists^\infty$. Note that now we can reformulate the "standard" quantifiers $\exists, \forall$ using modulo counting quantifiers, simply by replacing $\forall x \varphi(x)$ by $\exists^=0 (\bmod \infty) x \neg \varphi(x)$ and $\exists x \varphi(x)$ by $\exists^\geq1 (\bmod \infty) x \varphi(x)$.

Using such interpretation allows us to obtain a shorter normal form for our logic, which makes the presentation reasonably simpler.

**Definition 1.** We say that a formula $\varphi \in \text{FO}^1_{\text{MOD}}$ is flat, if:

$$ \varphi = \bigwedge_{i=1}^n \exists^{=a_i (\bmod b_i)} x \psi_i(x), $$

where $\leq \in \{\leq, \geq\}$, each $a_i$ is a natural number, each $b_i$ is a natural number or infinity and all $\psi_i$ are quantifier-free formulas.

The main purpose of introducing a flat form for $\text{FO}^1_{\text{MOD}}$–formulas is to avoid nested quantification and to concentrate only on a single element of a model. Furthermore, flattening makes the proofs and algorithm simpler. The following lemma shows that every satisfiable $\text{FO}^1_{\text{MOD}}$–formula can be flattened in \textit{NPTime}:
Lemma 1. There exists a nondeterministic polynomial time procedure, taking as its input an \( \text{FO}^1_{\text{MOD}} \)–formula over a signature \( \tau \) and producing a flat formula \( \varphi' \) over the same signature \( \tau \), such that \( \varphi \) is satisfiable iff the procedure has a run producing a satisfiable \( \varphi' \).

Proof. Lemma can be proved in a standard fashion, similarly to Theorem 1 in [12].

Fix a signature \( \tau \), and following a standard terminology, an atomic 1-type over \( \tau \) is a maximal satisfiable set of atoms or negated atoms involving only the variable \( x \). Usually we identify a 1-type with the conjunction of all its elements. Note that the number of all atomic 1-types is exponential in the size of \( \tau \).

2.2 Linear algebra

A linear inequality is an expression of the form \( t \geq t' \), where \( t \) and \( t' \) are linear terms. In this paper we are interested only in linear inequalities with integer coefficients. It is a well known that solving such systems of inequalities is in \( \text{NPTime} \) [4]. We recall some classical results concerning systems of linear inequalities. The first one gives us an estimation on the solution size:

Lemma 2 ([10]). Let \( \mathcal{E} \) be a system of \( I \) inequalities with \( U \) unknowns. Assume that all coefficients are integers absolutely bounded by \( C \). If there is a solution for the system \( \mathcal{E} \) over \( \mathbb{N} \), there is also a solution in which the values assigned to the unknowns are all bounded by \( U(1C)^{2I+1} \).

The second lemma provides an upper bound on the number of non-zero unknowns in a solution:

Lemma 3 ([6], Theorem 2). Let \( \mathcal{E} \) be a system of \( I \) inequalities with integer coefficients such that the absolute value of each coefficient from \( \mathcal{E} \) is bounded by \( C \). If \( \mathcal{E} \) has a solution over \( \mathbb{N} \), then it has a solution over \( \mathbb{N} \) with the number of non-zero unknowns bounded by \( 2I \log (4IC) \).

3 The finite satisfiability of \( \text{FO}^1_{\text{MOD}} \)

In this section we will show that one-variable fragment of first-order logic remains \( \text{NPTime} \)-complete even if we extend it with modulo counting quantifiers. We are interested only in the finite models, since our modulo quantifiers do not make sense over infinite structures. Our proof will strongly rely on techniques presented in [12], namely reducing our problem to solving some system of linear inequalities and integer programming.

Let \( \varphi \) be a satisfiable \( \text{FO}^1_{\text{MOD}} \)–formula over a signature \( \tau \) and let \( \mathfrak{M} \) be its finite model. Due to Lemma 1, we can assume that \( \varphi \) is flat. Let \( \pi_1, \pi_2, \ldots, \pi_{2|\tau|} \) be all possible 1-types over the signature \( \tau \).

We will first sketch our approach. Since the expressive power of \( \text{FO}^1_{\text{MOD}} \) is very limited, using this logic we can only describe properties of a single element
in a model and write sentences about the total number of elements satisfying some properties. Thus any model of $\varphi$ can be described, up to isomorphism, by a characteristic vector $\chi_\varphi = (\# \pi_1, \# \pi_2, \ldots, \# \pi_{2^{|r|}}) \in \mathbb{N}^{2^{|r|}}$, where $\# \pi_i$ is the number of occurrences of a 1-type $\pi_i$ in the model. Our main goal is to translate our formula $\varphi$ into a system of inequalities and congruences $\mathcal{E}$, which solution will be a tuple $\chi_\varphi$. Finally, we will replace all congruences by introducing some fresh variables and adding some new inequalities to the system $\mathcal{E}$. Solving systems of inequalities is in NPTIME, so it gives us ExpTIME solution, since we have exponentially many unknowns in $\mathcal{E}$. To improve it, we will use Lemma 3, which states that if there is a solution for $\mathcal{E}$, there is also a solution with only polynomially many non-zero values.

We will focus on transforming the formula $\varphi$ into a system of inequalities. Fix $1 \leq j \leq n$ and let $\varphi_j = \exists^{\pi_1, a_j}(\text{mod } b_j) x \psi_j(x)$ be the $j$-th conjunct from $\varphi$.

We define an indicator $\mathbb{1}^{(j)} = \begin{cases} 1, & \text{if } \pi_i \to \psi_j \text{, which intuitively tells us if} \\ 0, & \text{otherwise} \end{cases}$, which multiplies the number of occurrences of a 1-type, i.e. $S_j = \sum_{i=1}^{2^{|r|}} \mathbb{1}^{(j)} \cdot \# \pi_i$. That gives us the following system of congruences:

$$\mathcal{E}_0 = \begin{cases} \mathbb{1}^{(1)}_1 x_1 + \mathbb{1}^{(1)}_2 x_2 + \ldots + \mathbb{1}^{(1)}_{2^{|r|}} x_{2^{|r|}} \mod b_1 \bowtie_1 a_1 \\ \vdots \\ \mathbb{1}^{(n)}_1 x_1 + \mathbb{1}^{(n)}_2 x_2 + \ldots + \mathbb{1}^{(n)}_{2^{|r|}} x_{2^{|r|}} \mod b_n \bowtie_n a_n \end{cases}$$

Now we are going to replace congruences inside $\mathcal{E}_0$ with inequalities. We consider two cases, when $b_j$ is the infinity and when $b_j$ has a finite value. In the former case we can simply replace the congruence $(S_j \mod \infty) \bowtie_j a_j$ with $S_j \bowtie_j a_j$, since mod $\infty$ is just an abbreviation of a standard counting quantifier. Let us concentrate on the latter case. Observe that for any number $S_j$, there exists a reminder $r_j \in \mathbb{Z}_{b_j}$ and a quotient $q_j \in \mathbb{N}$, such that $S_j = r_j + q_j b_j$. Thus we only need to ensure that the reminder satisfies the inequality $r_j \bowtie_j a_j$. However $r_j$ and $q_j$ could not be constant, so we need to introduce them as two extra variables (respectively $y_j$ and $z_j$). We can rewrite our system $\mathcal{E}_0$ in the following way:

$$\mathcal{E} = \begin{cases} \mathbb{1}^{(1)}_1 x_1 + \mathbb{1}^{(1)}_2 x_2 + \ldots + \mathbb{1}^{(1)}_{2^{|r|}} x_{2^{|r|}} = y_1 + z_1 \cdot b_1 \\ y_1 \bowtie_1 a_1 \\ y_1 \leq b_1 - 1 \\ \vdots \\ \mathbb{1}^{(n)}_1 x_1 + \mathbb{1}^{(n)}_2 x_2 + \ldots + \mathbb{1}^{(n)}_{2^{|r|}} x_{2^{|r|}} = y_n + z_n \cdot b_n \\ y_n \bowtie_n a_n \\ y_n \leq b_n - 1 \end{cases}$$
Some cosmetic changes are left – the equality symbol is not allowed in \( \mathcal{E} \), so we need to replace it with \( \leq \) and \( \geq \) symbols.

Let \( \mathcal{E} \) be the system of inequalities obtained by the procedure described above. Observe that the number of inequalities in the system \( \mathcal{E} \) is bounded by \( 4n \in O(||\varphi||) \), the number of unknowns is bounded by \( 2^{|\varphi|} + 2n \in O(2^{||\varphi||}) \) and the absolute value of each coefficient is bounded by \( \max_i \{ \{a_i, b_i \} \setminus \{ \infty \} \} \in O(2^{||\varphi||}) \) (since we assume that \( a_i, b_i \) are encoded in binary system).

Using Lemma 3, we know that the number of non-zero unknowns in \( \mathcal{E} \) can be bounded by some polynomial function of \( ||\varphi|| \). By this fact, we can simply guess which unknowns will have non-zero values and write the system of inequalities only for them. Then the system has polynomial size in \( ||\varphi|| \), thus could be solved in NPTIME.

Below we present a nondeterministic polynomial time algorithm for checking if a given \( \text{FO}^1_{\text{MOD}} \)-formula has a finite model.

---

**Procedure 1** \( \text{FO}^1_{\text{MOD}} \text{-sat-test} \)

**Input:** a \( \text{FO}^1_{\text{MOD}} \)-formula \( \varphi \)

1. **Guess** \( \varphi' \) – a flattened \( \varphi \).
2. **Guess** which 1-types are realized at least one time.
3. Write the system of inequalities \( \mathcal{E} \) for the guessed 1-types.
4. Return **True**, if \( \mathcal{E} \) has a solution over \( \mathbb{N} \) and **False** otherwise.

---

**Lemma 4.** A formula \( \varphi \in \text{FO}^1_{\text{MOD}} \) has a finite model if and only if Procedure 1 returns **True**.

**Proof.** Assume that \( \varphi \) has a finite model. Therefore, we can obtain a flat satisfiable formula \( \varphi' \) from \( \varphi \) and describe its model with the system of linear inequalities \( \mathcal{E} \). Clearly the system \( \mathcal{E} \) has a solution over \( \mathbb{N} \) (e.g. \( \chi_{\varphi'} \)), hence Procedure 1 accepts \( \varphi \).

Suppose that Procedure 1 accepts the formula \( \varphi \). Then its model can be easily constructed, simply by taking a proper number of elements with a given 1-type, exactly as described in the solution of the constructed system of linear inequalities.

**Theorem 1.** The satisfiability problem for \( \text{FO}^1_{\text{MOD}} \) is NPTIME-complete.

**Proof.** The lower bound comes from boolean satisfiability problem, which can be easily encoded in \( \text{FO}^1 \) by replacing each propositional variable \( p \) in a boolean formula \( \varphi \) with a unary atom \( p(x) \) and prepending thus obtained formula \( \varphi(x) \) by quantifier \( \exists x (\text{mod } \infty) \).

11
For the upper bound, note that Procedure 1 works in \text{NPTime}, since flattening process can be done in \text{NPTime}, the maximum number of realized 1-types in model is bounded polynomially in $\|\varphi\|$ and solving systems of inequalities with polynomially many unknowns is in \text{NPTime}.

Another interesting property of the presented algorithm is its optimality:

**Lemma 5.** The size of models constructed by Procedure 1 is essentially optimal.

**Proof.** At the beginning, note that it is possible to enforce that each model of a formula $\varphi$ has an exponential size in $\|\varphi\|$. It can be done simply by considering a formula $\varphi = \exists = 2^n \text{ (mod } \infty)$, which every model has $2^n \in \mathcal{O}(2^{\|\varphi\|})$ elements.

Consider the system of inequalities $\mathcal{E}$ constructed in Procedure 1. Recall that $I$ is the number of inequalities in $\mathcal{E}$, $U$ is the number of unknowns and $C$ is a maximum of absolute values of coefficients. Observe that Lemma 2 gives us an $U(I)(IC)^{2I+1}$ upper bound on the maximal value of each unknown and Lemma 3 tells us that we can consider only the systems with at most $2I \log (4IC)$ unknowns. Thus the size of a constructed model, described by the solution of the system $\mathcal{E}$, is bounded by $2I \log (4IC) \cdot (IC)^{2I+1}$. After some trivial simplifications the obtained bound is clearly exponential in $\|\varphi\|$.

4 Future and Ongoing work

We proved that the finite satisfiability problem for \(\text{FO}^1_{\text{MOD}}\) is \text{NPTime}-complete. The next natural question is what happens in the case of the two-variable fragment. Over arbitrary structures we do not know the answer at this moment. As we mentioned in Introduction it is known that \(\text{FO}^2\) over words is \text{ExpSpace}-complete i.e. it is harder than \(\text{FO}^2\) extended with the standard counting quantifiers on words which is \text{NExpTime}-complete [5]. Currently we are working on the case of finite (ordered, unranked) trees. Let us recall that with the standard counting quantifiers \(\text{FO}^2\) over trees is \text{ExpSpace}-complete [2]. As in the case of finite words, modulo counting quantifiers allow to lift the lower bound, in this case to \(2\text{ExpTime}\). The proof goes by a reduction from alternating Turing machines with exponentially bounded space and is a natural adaptation of the lower bound proof from [15]. We also know how to show a matching upper bound, which is much more involved. Generally the proof extends the construction for the case of \(\text{FO}^2\) with the standard counting quantifiers from [2].

References

Towards Symbolic Factual Change in DEL

Malvin Gattinger

Institute for Logic, Language & Computation, University of Amsterdam

Abstract. We extend symbolic model checking for Dynamic Epistemic Logic (DEL) with factual change. Our transformers provide a compact representation of action models with pre- and postconditions, for both S5 and the general case. The method can be implemented using binary decision diagrams and we expect it to improve model checking performance. As an example we give a symbolic representation of the Sally-Anne false belief task.

Keywords: Epistemic Logic, Symbolic Model Checking, Factual Change

1 Introduction

Symbolic representation is a solution to the state explosion problem in model checking. The idea is to not store models explicitly in memory, but to find more compact representations which still allow the evaluation of formulas. In [3] it was shown that S5 Kripke models for Dynamic Epistemic Logic (DEL) can be encoded symbolically using knowledge structures. They are of the form $(V, \theta, O)$ where $V$ is a set of atomic propositions called vocabulary, $\theta$ is a boolean formula called state law and $O_i \subseteq V$ are observational variables for each agent. Notably, this symbolic representation preserves the truth of all DEL formulas, including higher-order knowledge (see Section 3 below).

The framework was generalized in [4] in two ways: From equivalences to arbitrary relations and from announcements to action models. The latter can be represented by knowledge transformers of the form $(V^+, \theta^+, O^+)$ analogous to the product update on Kripke models [1], applying a transformer to a structure yields a new structure. However, knowledge transformers only change what agents know and not what is the case — they do not provide a symbolic equivalent of postconditions for factual change as studied in [5].

In this paper we combine the two generalizations and add the missing components to treat factual change. The result are belief transformers with factual change which for simplicity we will just call transformers.

Possible worlds in a Kripke model get their meaning but not their identity via a valuation function. In particular we can assign the same atomic truths to different possible worlds. In contrast, all states of a knowledge structure satisfy different atomic propositions and can thus be identified with their valuation. This is what makes structures symbolic and efficient to implement, but it complicates the idea of changing facts, as the following minimal example shows.
Example 1. Consider a coin lying on a table with heads up; \( p \) is true and this is common knowledge. Suppose we then toss it randomly and hide the result from agent \( a \) but reveal it to agent \( b \). Figure 1 shows a Kripke model of this update.

It is easy to find the following structures that are equivalent to the initial and the resulting model, but how can we symbolically describe the update which transforms one into the other?

\[
\begin{align*}
(V = \{p\}, \theta = p, O_a = \{p\}, O_b = \{p\}) \\
\times \\
(V = \{p\}, \theta = \top, O_a = \emptyset, O_b = \{p\})
\end{align*}
\]

The name of a resulting world \((w, a_1)\) makes clear that it “comes from” \( w \). But a state like \( \emptyset \) does not reveal its history or any relation to \( \{p\} \). For purely epistemic actions this is not a problem — we only add propositions from \( V^+ \) to the state to distinguish different epistemic events. But for factual change propositions from \( V \) have to be modified and we need a way to remove them from states.

Our solution is to copy propositions: We store the old value of \( p \) in a fresh variable \( p^r \). Then we rewrite the state law and observations using substitutions.

We proceed as follows. Sections 2 and 3 summarize the relevant parts of [4], generalized to belief transformers. We then add factual change in Section 4 and show that transformers are equivalent to action models in Section 5. The Sally-Anne task illustrates our framework in Section 6 and we finish with further questions in Section 7.

Definition 1 (Languages and Notation). We fix a finite set of agents \( I \) denoted by \( i, j, \) etc. and use the letters \( V \) or \( X \) for sets of atomic propositional variables denoted by \( p, q, \) etc.

For any set of propositions \( X \) we write \( \mathcal{L}_B(X) \) for the boolean language given by the BNF \( \varphi ::= p \mid \neg \varphi \mid \varphi \land \varphi \) where \( p \in X \). Similarly, let \( \mathcal{L}(X) \) be the epistemic language over \( X \) given by \( \varphi ::= p \mid \neg \varphi \mid \varphi \land \varphi \mid \Box_i \varphi \) where \( i \in I \).
Primes and circles denote fresh variables, for example $p'$ and $p^\circ$. For sets of variables let $X' := \{ x' \mid x \in X \}$ and $X^\circ := \{ x^\circ \mid x \in X \}$. We also extend this notation to formulas, for example $(\Box p \land q') = (\Box(p \land q')$.

We write $[p/\psi]\varphi$ for the result of substituting $\psi$ for $p$ in $\varphi$. Given two sets of the same size $A$ and $B$ of atomic propositions, and implicitly assuming an enumeration $A = \{a_1, \ldots, a_k\}$ and $B = \{b_1, \ldots, b_k\}$ we write $[B/A]\varphi$ for the result of substituting $a_i$ for $b_i$ in $\varphi$ in parallel for all $i$.

A boolean assignment is identified with its set of true propositional variables and we write $\models$ for the standard satisfaction relation. Boolean quantification is used as follows: $\forall p \varphi := [p/\top] \varphi \land [p/\bot] \varphi$. For any $A = \{p_1, \ldots, p_n\}$, let $\forall A \varphi := \forall p_1 \forall p_2 \ldots \forall p_n \varphi$. To abbreviate that a specific subset of propositions is true, let $A \subseteq B$.

Definition 1 describes operations on boolean formulas which might not be efficient in practice. In any actual implementation of our methods those should be replaced with operations on boolean functions represented as Binary Decision Diagrams (BDDs) [7]. For example, boolean quantification should not be implemented as an abbreviation but can instead be done efficiently by eliminating quantified variables from the BDD.

## 2 Kripke Models and Action Models

We quickly state the standard definitions for Kripke semantics of DEL. For a general introduction see [9] and for details on factual change see [5].

**Definition 2.** A Kripke model for $V$ is a tuple $\mathcal{M} = (W, R, \pi)$ where $W$ is the set of worlds, $R_i \subseteq W \times W$ is a relation for each $i$ and $\pi : W \rightarrow P(V)$ is a valuation function. A pointed Kripke model is a pair $(\mathcal{M}, w)$ where $w \in W$.

We interpret $L(V)$ on pointed Kripke models as follows.

1. $(\mathcal{M}, w) \models p$ iff $p \in \pi(w)$
2. $(\mathcal{M}, w) \models \neg \varphi$ iff not $(\mathcal{M}, w) \models \varphi$
3. $(\mathcal{M}, w) \models \varphi \land \psi$ iff $(\mathcal{M}, w) \models \varphi$ and $(\mathcal{M}, w) \models \psi$
4. $(\mathcal{M}, w) \models \Box \varphi$ iff for all $v \in W$ : If $wR_i v$ then $(\mathcal{M}, v) \models \varphi$

The following definition describes action models and how they can be applied to Kripke models. Our definition of postconditions differs from the standard in [5] because we only allow boolean formulas. This however does not change the expressivity [10].

**Definition 3.** An action model is a tuple $A = (A, R, \text{pre}, \text{post})$ where $A$ is a set of atomic events, $R_i \subseteq A \times A$ a relation for each $i$, $\text{pre} : A \rightarrow L(V)$ is a precondition function and $\text{post} : A \times V \rightarrow L_B(V)$ a postcondition function.

The product update is defined by $\mathcal{M} \times A := (W^{\text{new}}, R_i^{\text{new}}, \pi^{\text{new}})$ where

- $W^{\text{new}} := \{ (w, a) \in W \times A \mid \mathcal{M}, w \models \text{pre}(a) \}$
- $R_i^{\text{new}} := \{ ((w, a), (v, b)) \mid R_i w v \text{ and } R_i b \}$
- $\pi^{\text{new}}((w, a)) := \{ p \in V \mid \mathcal{M}, w \models \text{post}_i(p) \}$

An action is a pair $(A, a)$ where $a \in A$. 

16
3 Belief Structures and Belief Transformers

We now present the definitions of belief structures and belief transformers. The key idea is that instead of explicitly listing worlds we use a symbolic representation: The set of worlds and the valuation function are replaced by a vocabulary and a boolean formula called the state law. The set of states is then implicitly given as all boolean assignments that satisfy the state law, i.e. a subset of the powerset of the vocabulary. Moreover, also the epistemic relations can be encoded using boolean formulas and the goal is to interpret the language on the resulting structures without ever computing or listing the full set of states. For more details and proofs, we refer the reader to [4].

Definition 4. A belief structure is a tuple \( \mathcal{F} = (V, \theta, \Omega) \) where \( V \) is a finite set of atomic propositions called vocabulary, \( \theta \in \mathcal{L}_B(V) \) is the state law and \( \Omega_i \in \mathcal{L}_B(V \cup V^+) \) are called observations. Any \( s \subseteq V \) such that \( s \models \theta \) is called a state of \( \mathcal{F} \). A pair \( (\mathcal{F}, s) \) where \( s \) is a state of \( \mathcal{F} \) is called a scene.

We interpret \( \mathcal{L}(V) \) on scenes as follows.

1. \((\mathcal{F}, s) \models p \) iff \( s \models p \).
2. \((\mathcal{F}, s) \models \neg \varphi \) iff not \((\mathcal{F}, s) \models \varphi \).
3. \((\mathcal{F}, s) \models \varphi \land \psi \) iff \((\mathcal{F}, s) \models \varphi \) and \((\mathcal{F}, s) \models \psi \).
4. \((\mathcal{F}, s) \models \Box_i \varphi \) iff for all \( t \subseteq V : \text{If } t \models \theta \text{ and } (s \cup t') \models \Omega_i \text{ then } (\mathcal{F}, t) \models \varphi \).

We write \((\mathcal{F}, s) \models_V (\mathcal{F}', s')\) iff these two scenes agree on all formulas of \( \mathcal{L}(V) \).

An interesting property of belief structures is that on a given structure all epistemic formulas have boolean equivalents. The following translation reduces model checking to boolean operations which is not possible on Kripke models.

Definition 5. For any belief structure \( \mathcal{F} = (V, \theta, \Omega) \) and any formula \( \varphi \in \mathcal{L}(V) \) we define its local boolean translation \( \| \varphi \|_{\mathcal{F}} \) as follows.

1. For any primitive formula, let \( \| p \|_{\mathcal{F}} : = p \).
2. For negation, let \( \| \neg \varphi \|_{\mathcal{F}} : = \neg \| \varphi \|_{\mathcal{F}} \).
3. For conjunction, let \( \| \psi_1 \land \psi_2 \|_{\mathcal{F}} : = \| \psi_1 \|_{\mathcal{F}} \land \| \psi_2 \|_{\mathcal{F}} \).
4. For belief, let \( \| \Box_i \psi \|_{\mathcal{F}} : = \forall V' (\theta' \rightarrow (\Omega_i \rightarrow (\| \psi \|_{\mathcal{F}}))) \).

Theorem 1 (from [4]). Definition 5 preserves and reflects truth. That is, for any formula \( \varphi \) and any scene \((\mathcal{F}, s)\) we have that \((\mathcal{F}, s) \models \varphi \) iff \( s \models \| \varphi \|_{\mathcal{F}} \).

The following definition was only hinted at in [4]. Belief transformers are like knowledge transformers, but instead of observed propositions \( O_i^+ \) we use boolean formulas \( \Omega_i^+ \) to encode arbitrary relations on \( \mathcal{P}(V^+) \).

Definition 6. A belief transformer for \( V \) is a tuple \( \mathcal{X} = (V^+, \theta^+, \Omega^+) \) where \( V^+ \) is a set of atomic propositions such that \( V \cap V^+ = \emptyset \) and \( \theta^+ \in \mathcal{L}(V \cup V^+) \) is a possibly epistemic formula and \( \Omega_i^+ \in \mathcal{L}_B(V \cup V^+) \) is a boolean formula for each \( i \in I \). A belief event is a belief transformer together with a subset \( x \subseteq V^+ \), written as \((\mathcal{X}, x)\).
The belief transformation of a belief structure $F = (V, \theta, \Omega)$ with $X$ is defined by $F \times X := (V \cup V^+, \theta \land ||\theta^+||_F, \{\Omega_i \land \Omega_i^+\}_{i \in I})$. Given a scene $(F, s)$ and a belief event $(X, x)$, let $(F, s) \times (X, x) := (F \times X, s \cup x)$.

The resulting observations are boolean formulas over the new double vocabulary $(V \cup V') \cup (V^+ \cup V'^+)$, describing a relation between the new states which are subsets of $V \cup V^+$.

4 Belief Transformers with Factual Change

We now define transformation with factual change, adding the components $V_-$ and $\theta_-$. Note that the belief transformers without factual change as discussed in the previous section are exactly those transformers where $V_- = \emptyset$.

Definition 7. A transformer for $V$ is a tuple $X = (V^+, \theta^+, V_-, \theta_-, \Omega^+)$ where

- $V^+$ is a set of fresh atomic propositions such that $V \cap V^+ = \emptyset$,
- $\theta^+$ is a possibly epistemic formula from $\mathcal{L}(V \cup V^+)$,
- $V_- : V_- \to \mathcal{L}_B(V \cup V^+)$ is the modified subset of the original vocabulary,
- $\theta_- : V_- \to \mathcal{L}_B(V \cup V^+)$ maps modified propositions to boolean formulas,
- $\Omega_+^i \in \mathcal{L}_B(V^+ \cup V'^+)$ are boolean formulas for each $i \in I$.

To transform $F = (V, \theta, \Omega_i)$ with $X$, let $F \times X := (V^{new}, \theta^{new}, \Omega_i^{new})$ where

1. $V^{new} := V \cup V^+ \cup V_-$
2. $\theta^{new} := [V_-/V^+] (\theta \land ||\theta^+||_F) \land \bigwedge_{q \in V_-} (q \leftrightarrow [V_-/V^+] (\theta_- (q)))$
3. $\Omega^{new}_i := ([V_-/V^+] (\theta_- ) \land \bigwedge_{q \in V_-} (q \leftrightarrow [V_-/V^+] (\theta_- (q)))) \land \Omega_i^+$

An event is a pair $(X, x)$ where $x \subseteq V^+$. Given $(F, s)$ and $(X, x)$, let $(F, s) \times (X, x) := (F \times X, s^{new})$ where the new actual state is $s^{new} := (s \setminus V_-) \cup (s \cap V_-)^p \cup x \cup \{p \in V_- | s \cup x \vdash \theta_- (p)\}$.

To explain this definition, let us consider the components one by one.

First, the new vocabulary contains $V_- = \{p^\circ | p \in V_-\}$. These are fresh copies of the modified subset. We use them to keep track of the old valuation.

Second, the new state law: A state in the resulting structure needs to satisfy the old state law and the event law encoding the preconditions. For modified propositions the old values have to be used, hence we apply a substitution to both laws in the left conjunct. Modified propositions are then overwritten in the right conjunct, using $\theta_-$ which encodes postconditions. As in Definition 3, postconditions are evaluated in the old model, hence we also substitute here.

Third, for the new observations we replace modified variables by their copies. Two substitutions are needed because $\Omega_i^{new}$ is in a double vocabulary. Old observations induce new ones via the state law. For example, if $q$ was flipped publicly, then $q \leftrightarrow \neg q^\circ$ is part of the new state law and observing whether $q$ is equivalent to observing whether $\neg q^\circ$, i.e. having observed $q$ in the original structure. In the simpler $S5$ setting we would use $\Omega_i^{new} := ([V_-/V^+] O_i) \cup O_i^+$. 
Finally, the new actual state \( s^{\text{new}} \) is the union of, in this order: propositions in the old state that have not been modified \((s \setminus V_-)\), copies of the modified propositions that were in the old state \((s \cap V_-)\), those propositions labeling the actual event \(x\) and the modified propositions whose precondition was true in the old state \(\{p \in V_- \mid s \cup x \models \theta_-(p)\}\).

**Example 2.** We can now model the coin flip from Example 1 as follows. Because we use the more general belief (instead of knowledge) structures, the initial structure now has boolean formulas \(\Omega_i\) instead of observational variables \(O_i\):

\[(V = \{p\}, \theta = p, \Omega_a = p \leftrightarrow p', \Omega_b = p \leftrightarrow p')\]

The following transformer models the coin flip visible to \(b\) but not to \(a\):

\[(V^+ = \{q\}, \theta^+ = \top, V_- = \{p\}, \theta_-(p) := q, \Omega_a^+ = \top, \Omega_b^+ = q \leftrightarrow q')\]

The result of applying the latter to the former is this:

\[(V = \{p, q, p^\circ\}, \theta = p^\circ \land (p \leftrightarrow q), \Omega_a = p^\circ \leftrightarrow p^\circ', \Omega_b = (p^\circ \leftrightarrow p^\circ') \land (q \leftrightarrow q'))\]

**Example 3.** A publicly observable change \(p := \varphi\) for a propositional formula \(\varphi\) is modeled by:

\[(V^+ = \emptyset, \theta^+ = \top, V_- = \{p\}, \theta_-(p) := \varphi, \Omega_i^+ = \top)\]

DEL does not have temporal operators and agents never know the past explicitly. Hence the old valuation is often irrelevant and the product update on Kripke models does this “garbage collection” better than our transformation. But we can eliminate propositions outside the original \(V\) using the following Lemma. A more thorough analysis of minimizing knowledge structures will be future work.

**Lemma 1.** Suppose \(F\) uses the vocabulary \(V \cup \{p\}\) and \(p \notin V\) is determined by the state law (i.e. \(\theta \rightarrow p\) or \(\theta \rightarrow \neg p\) is a tautology). Then we can remove \(p\) from the state law and observational BDDs to get a new structure \(F'\) using the vocabulary \(V\) such that \((F, s) \equiv_V (F', s \setminus \{p\})\).

**Example 4.** The result from Example 2 is \(\equiv_{(p,q)}\) equivalent to:

\[(V = \{p, q\}, \theta = p \leftrightarrow q, \Omega_a = \top, \Omega_b = q \leftrightarrow q')\]

## 5 Equivalence and Expressiveness

We now show that transformers describe exactly the same class of updates as action models. The main ingredients for the proof are the following Lemma and two Definitions of how to go from transformers to action models and back.

**Lemma 2 (from [4]).** Suppose we have a belief structure \(F = (V, \theta, \Omega)\), a finite Kripke model \(M = (W, \pi, \mathcal{R})\) for the vocabulary \(X \subseteq V\) and a function \(g : W \rightarrow \mathcal{P}(V)\) such that
Then, for every $\mathcal{L}(X)$-formula $\varphi$ we have $(\mathcal{F}, g(w)) \models \varphi$ iff $(M, w) \models \varphi$.

**Definition 8 (Act).** Given an event $(X = (V^+, \theta^+, V_-, \theta_-, \Omega^+), x)$, define an action $(\text{Act}(X) := (A, \text{pre}, \text{post}, R), a := x)$ by

- $A := \mathcal{P}(V^+)$
- $\text{pre}(a) := \lceil a/\top \rceil [\lceil V^+ \setminus a \rceil/\bot] \theta^+$
- $\text{post}(p) := \{ p \}$
- $R := \{(a, b) \mid a \cup (b') \models \Omega^+ \}$

**Definition 9 (Trf).** Consider an action $(A = (\text{pre}, \text{post}, R), a_0)$. Let $n := \text{ceil}(\log_2|A|)$ and $\ell : A \to \mathcal{P}\{q_1, \ldots, q_n\}$ be an injective labeling function using fresh atomic variables $q_i$. Then let $(\text{Trf}(A) := (V^+, \theta^+, V_-, \theta_-, \Omega^+), x := \ell(a_0))$ be the event defined by

- $V^+ := \{q_1, \ldots, q_n\}$
- $\theta^+ := \bigvee_{a \in A} (\text{pre}(a) \land \ell(a) \subseteq V^*)$
- $V_- := \{ p \in V \mid \exists a : \text{post}(p) \neq p \}$
- $\theta_-(p) := \bigvee_{a \in A} (\ell(a) \subseteq V^+ \land \text{post}(p))$
- $\Omega^+_\ell := \bigvee_{(a,b) \in R} (a \subseteq V^+ \land (\ell \in \Omega^+)')$

Besides these translations for the dynamic parts, we also use the translations $\mathcal{M}(\cdot)$ and $\mathcal{F}(\cdot)$ from structures to models and vice versa, as given in Definitions 18 and 19 of [4]. Now everything is in place to state and prove our main result. The following generalizes Theorem 4 in [4].

**Theorem 2.** (i) Definition 8 is truthful: For any scene $(\mathcal{F}, s)$, any event $(X, x)$ and any formula $\varphi$ over the vocabulary of $F$ we have:

$$(\mathcal{F}, s) \times (X, x) \models \varphi \iff (\mathcal{M}(\mathcal{F}), s) \times (\text{Act}(X), x) \models \varphi$$

(ii) Definition 9 is truthful: For any pointed Kripke model $(M, w)$, any action $(A, a)$ and any formula $\varphi$ over the vocabulary of $M$ we have:

$$(M \times A, (w, a)) \models \varphi \iff (\mathcal{F}(M), g_M(w)) \times (\text{Trf}(A), \ell(a)) \models \varphi$$

where $g_M$ is from $\mathcal{F}(M)$ in Definition 19 of [4] and $\text{Trf}(A)$ and $\ell$ are from Definition 9 above.

**Proof.** By Lemma 2. We first need appropriate functions $g$.

For part (i), $g$ needs to map worlds of $\mathcal{M}(\mathcal{F}) \times \text{Act}(X)$, i.e. pairs $(s, x) \in \mathcal{P}(V) \times \mathcal{P}(V^+)$ to states of $\mathcal{F} \times X$, i.e. subsets of $V \cup V^+ \cup V_\bot$. Let $g(s, x) := (s \setminus V_-) \cup (s \setminus V^+) \cup x \cup \{ p \in V_- \mid s \cup x \models \theta_-(p) \}$ which is exactly $s^{\text{new}}$ from Definition 7 above. We now prove C1 to C3 from Lemma 2.
For C1, take any two worlds \((s, x)\) and \((t, y)\). We need to show \(g(s, x)(g(t, y))^\prime \models \Omega_i^{\text{new}}\) iff \(\mathcal{R}^{\text{new}}_i(s, x)(t, y)\). For this, note the following equivalences. We have \(g(s, x)(g(t, y))^\prime \models \Omega_i^{\text{new}}\) iff
\[
(s \setminus V_-) \cup (s \cap V_-)^\circ \cup x \cup \{ p \in V_- \mid s \cup x \vDash \theta_-(p) \}
\cup \{(t \cap V_-)^\circ \cup y \cup \{ p \in V_- \mid t \cup y \vDash \theta_-(p) \} \}
\models \left[ V_- / V_-^2 \right] \left[ (V_-') / (V_-') \right] \Omega_i \land \Omega_i^\dagger
\]
Here \(V_-\) and \(V_-^\dagger\) do not occur in the formula, as old epistemic relations do not depend on new values of modified propositions. Hence we can drop the subsets of \(V_-\) and \(V_-^\dagger\) to obtain the equivalent condition
\[
(s \setminus V_-) \cup (s \cap V_-)^\circ \cup x \cup \{ p \in V_- \mid s \cup x \vDash \theta_-(p) \}
\cup \{(t \cap V_-)^\circ \cup y \cup \{ p \in V_- \mid t \cup y \vDash \theta_-(p) \} \}
\models \left[ V_- / V_-^2 \right] \left[ (V_-') / (V_-') \right] \Omega_i \land \Omega_i^\dagger
\]
in which we can split both sides into separate vocabularies:
\[
(s \setminus V_-) \cup (s \cap V_-)^\circ \cup x \cup \{ p \in V_- \mid s \cup x \vDash \theta_-(p) \}
\cup \{(t \cap V_-)^\circ \cup y \cup \{ p \in V_- \mid t \cup y \vDash \theta_-(p) \} \}
\models \left[ V_- / V_-^2 \right] \left[ (V_-') / (V_-') \right] \Omega_i \land \Omega_i^\dagger
\]
Now undo the \(\circ\)-substitution on both sides in the first conjunct to see that it is equivalent to \(s \cup t \vDash \Omega_i\). Hence the whole condition is equivalent to \(\mathcal{R}_{st}, \mathcal{R}_{tx} \) which is exactly \(\mathcal{R}^{\text{new}}(s, x)(t, y)\) by Definition of \(\mathcal{M}(\cdot)\) and Definition 8.

To show C2, take any \((s, x)\) and any \(p \in V\). We have to show that \(p \in g(s, x)\) iff \(p \in \pi^{\text{new}}(s, x) = \{ p \in V \mid \mathcal{M}, s \vDash \text{post}_x(p) \}\). There are two cases. First, if \(p \notin V_-\), then \(\text{post}_x(p) = p\) by Definition 8 and we directly have \(p \in g(s, x)\) iff \(p \in \pi^{\text{new}}(s, x)\). Second, if \(p \in V_-\), then \(p \in g(s, x)\) iff \(s \cup x \vDash \theta_-(p)\) by definition of \(g\) and \(\text{post}_x(p) = [x/\top] [(V^+ \setminus x) / \bot] \theta_-\). By Definition 8. Hence we have a chain of equivalences: \(p \in g(s, x)\) iff \(s \cup x \vDash \theta_-(p)\) iff \(s \vDash [x/\top] [(V^+ \setminus x) / \bot] \theta_-\) iff \(\mathcal{M}, s \vDash [x/\top] [(V^+ \setminus x) / \bot] \theta_-\) iff \(p \in \pi^{\text{new}}(s, x)\).

For C3, take any \(s^{\text{new}} \subseteq V \cup V^+ \cup V_-\). We want to show that \(s^{\text{new}} \models \theta_\circ \text{new}\) iff there is an \((s, x)\) such that \(g(s, x) = s^{\text{new}}\).

For left-to-right, suppose \(s^{\text{new}} \models \theta_\circ \text{new}\), i.e. \(s^{\text{new}} = [V_- / V_-^2] [\theta \land \| \theta^+ \| x] \land \bigwedge_{q \in V_-} (q \leftrightarrow [V_- / V_-^2] (\theta^-(q)))\). Now first, let \(s^{\text{old}} := (s^{\text{new}} \cap V_-) \cup \{ p \in V_- \mid p^\circ \in s^{\text{new}} \}\). We then have \(s^{\text{old}} \models \theta\), i.e. \(s^{\text{old}}\) is a state of \(\mathcal{F}\) and thus by the definition of \(\mathcal{M}(\cdot)\) also a world of \(\mathcal{M}(F)\). Second, let \(x := s^{\text{new}} \cap V^+\) and note that \(s \cup x \vDash \| \theta^+ \| x\). It can now be checked that \(g(s, x) = s^{\text{new}}\).

For right-to-left, suppose we have an \((s, x)\) such that \(g(s, x) = s^{\text{new}}\). Then we want to show \((s \setminus V_-) \cup (s \cap V_-)^\circ \cup x \cup \{ p \in V_- \mid s \cup x \vDash \theta_-(p) \} \models \left[ V_- / V_-^2 \right] [(\theta \land \| \theta^+ \| x] \land \bigwedge_{q \in V_-} (q \leftrightarrow [V_- / V_-^2] (\theta^-(q)))\) which indeed follows from \(s \models \theta\) and Definition 8.

For part (ii), \(q\) should map worlds of \(\mathcal{M} \times \mathcal{A}\) to states of \(\mathcal{F}(\mathcal{M}) \times \operatorname{Trf}(\mathcal{A})\). Again we use \(s^{\text{new}}\), but \(s\) and \(x\) are given by propositional encodings \(q_M(w)\) and \(\ell(a)\). Let \(g(w, a) := (g_M(w) \setminus V_-) \cup (g_M(w) \cap V_-)^\circ \cup \ell(a) \cup \{ p \in V_- \mid g_M(w) \cup \ell(a) \vDash \theta_-(p) \}\). We leave checking C1 to C3 as an exercise to the reader — the proofs are very similar to those in part (i). \(\square\)
6 Symbolic Sally-Anne

The Sally-Anne false belief task is a famous example used to illustrate and test for a theory of mind. The basic version goes as follows (adapted from [2]):

Sally has a basket, Anne has a box. Sally also has a marble and puts it in her basket. Then Sally goes out for a walk. Anne moves the marble from the basket into the Box. Now Sally comes back and wants to get her marble. Where will she look for it?

To answer this, one needs to realize that Sally did not observe that the marble was moved and will thus look for it in the basket. We now translate the first DEL modeling of this story from [6] to our framework. This choice is also motivated by a recent interest in the complexity of theory of mind [12,13] where our symbolic representation might provide a new perspective. For simplicity we adopt the naive modeling given in [6], leaving it as future work to also adopt the refinement with edge-conditions and other improvements of the model.

We use the vocabulary \( V = \{ p, t \} \) where \( p \) means that Sally is in the room and \( t \) that the marble is in the basket. In the initial scene Sally is in the room, the marble is not in the basket and both of this is common knowledge:

\[
(F_0, s_0) = ((V = \{ p, t \}, \theta = (p \land \neg t), \Omega_S = \top, \Omega_A = \top), \{ p \})
\]

The sequence of events is:

\( X_1 \): Sally puts the marble in the basket: \( ((\emptyset, \top, \{ p \}, \theta = (p \land \neg t), \Omega_S = \top, \Omega_A = \top), \emptyset) \).

\( X_2 \): Sally leaves: \( ((\emptyset, \top, \{ p \}, \theta = (p \land \neg t), \Omega_S = \top, \Omega_A = \top), \emptyset) \).

\( X_3 \): Anne puts the marble in the box, not observed by Sally: \( ((\emptyset, \top, \{ t \}, \theta = (p \land \neg t) \land (q \land \neg q'), q \leftrightarrow q'), \emptyset) \).

\( X_4 \): Sally comes back: \( ((\emptyset, \top, \{ p \}, \theta = (p \land \neg t), \Omega_S = \top, \Omega_A = \top), \emptyset) \).

We calculate the result in Figure 2, using Lemma 1 to remove superfluous variables. Note that all operations are boolean. Finally, we can check that in the last scene Sally believes the marble is in the basket:

\[
\{ p, q \} \models \Box \neg t
\]
\[
\iff \{ p, q \} \models \forall \theta' (\theta' \to (\Omega_S \to t'))
\]
\[
\iff \{ p, q \} \models \forall \theta' (\theta' \to \neg q' \land p' \to (\neg q' \to t'))
\]
\[
\iff \{ p, q \} \models \top
\]

7 Related and Future Work

We generalized knowledge transformers from [4] to belief transformers with factual change. The result is a new symbolic representation of action models with postconditions that can be implemented using binary decision decision diagrams [7].
The Sally-Anne false belief task is a famous example used to illustrate and test was moved and will thus look for it in the basket. We now translate the first DEL. The result is a new symbolic representation of action models with last scene Sally believes the marble is in the basket:

\[
\begin{align*}
\mathcal{F}_0 &= ((\{p, t\}, (p \land \neg t), T, T), p) \\
\mathcal{X}_1 &= ((\{p, t, t^\prime\}, (p \land \neg t^\prime) \land t, T, T), \{p, t\}) \\
\mathcal{X}_2 &= ((\{p, t\}, \theta_+(p) = \bot, T, T), \emptyset) \\
\mathcal{X}_3 &= ((\{p, t, t^\prime, p^\prime\}, (p^\prime \land \neg t^\prime) \land t \land \neg p, T, T), \{t, p^\prime\}) \\
\mathcal{X}_4 &= ((\{p, t\}, \theta_-(p) = T, T, T), \emptyset) \\
\mathcal{X}_5 &= ((\{p, t, t^\prime, p^\prime\}, (p^\prime \land \neg t^\prime) \land t \land \neg p, T, T), \{t, p^\prime\}) \\
\end{align*}
\]

Fig. 2. Sally-Anne on belief structures and transformers.

As mentioned above, restricting postconditions to boolean formulas does not limit the expressivity. The authors of [10] in fact prove the stronger result that postconditions can be restricted to $T$ and $\bot$. Hence one can also model postconditions as functions of the type $A \rightarrow P(V)$ as done in [6]. We leave it as future work to tune the definition of transformers in a similar way.

An alternative “succinct” representation for Kripke models and action models was recently developed in [8]. Succinct models also describe sets of worlds with not limit the expressivity. The authors of [10] in fact prove the stronger result that postconditions can be restricted to $T$ and $\bot$. Hence one can also model postconditions as functions of the type $A \rightarrow P(V)$ as done in [6]. We leave it as future work to tune the definition of transformers in a similar way.

Finally, the presented ideas are of course meant to be implemented. A natural next step therefore is to extend SMCDEL [11], the implementation of [4], with the presented transformers, working towards a symbolic model checker covering the whole Logic of Communication and Change from [5]. This work has been started and experimental modules including the Sally-Anne example are now available at https://github.com/jrclogic/SMCDEL.

Additionally, an implementation of [8] would be interesting to compare the performance of both approaches. Benchmark problems can be taken from both the DEL and the cognition literature, see for example [13].

Acknowledgements. Many thanks to Fernando R. Velázquez Quesada, Jan van Eijck and the anonymous reviewers for helpful comments and suggestions.
References

Certified Soundness of Simplest Known Formulation of First-Order Logic

John Bruntse Larsen

DTU Compute, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

Abstract. In 1965, Donald Monk published a paper about an axiomatic system for first-order predicate logic that he described as “the simplest known formulation of ordinary logic”. In this paper we show work in progress on certifying soundness of this system in the interactive proof assistant Isabelle. Through this work we demonstrate the usefulness of using proof assistants for validating mathematical results. This work also establishes an outline for future work such as a certified completeness proof of the axiomatic system in Isabelle.

Keywords: first-order logic, axiomatic system, soundness, proof assistant, Isabelle

1 Introduction

First-order predicate logic has a fundamental role in mathematics and computer science. It formalizes the concept of entities with properties and relations to other entities. It provides a framework for formal reasoning and proofs which are relevant in areas such as software engineering and AI. For this reason it is important that formalizations of first-order logic are correct so that they do not give erroneous results. By using an interactive proof assistant like Isabelle [2], we can work with formalizations in a certified manner as motivated by Geuvers [7] and Pfenning [6].

In this work we investigate using proof assistants for verifying a formulation of first-order logic with equality by J. Donald Monk [1]. In 1965, Monk published a paper about an axiomatic system for first-order predicate logic with equality that he described as “the simplest known formulation of ordinary logic” [3]. Simplicity in this context is to be understood as the simplicity of checking correctness of a proof. In the formulation, there are 10 axioms of which 2 of them have side conditions and the simplicity of the formulation is in particular the simplicity of checking these side conditions. To clarify the notion of simplicity in Monk’s formulation, we compare it with a textbook formulation of natural deduction and a formulation of sequent calculus.

The goal of this paper is to show work in progress on using the interactive proof assistant Isabelle to verify soundness of Monk’s formulation. The motivation for this work is to demonstrate the usefulness of working with a formulation from the 1960’s in a modern proof assistant. We begin with introducing some
of notions from [3–5] that are crucial for understanding Monk’s formulation and make comparisons with natural deduction and sequent calculus. We then describe how to verify soundness of the formulation in Isabelle where we use the following approach:

1. Defining the syntax of normalized first-order logic formulas.
2. Defining the semantics of the normalized formulas.
3. Defining the axioms and rules of Monk’s axiomatic system and a proof system for it.
4. Defining and proving soundness of the proof system.

By proving soundness of the proof system based on Monk’s axiomatic system, we obtain certified soundness of Monk’s formulation. We found the following Isabelle commands useful in making the formalization:

- **datatype** for defining a grammar in a BNF-style.
- **abbreviation** for defining abbreviations of terms such as \( \top \equiv \neg \bot \).
- **fun** for defining recursive functions.
- **definition** for defining non-recursive functions.
- **inductive** for defining inductive constructs.
- **lemma** for stating auxiliary lemmas for the soundness proof.
- **theorem** for stating soundness.

In the following sections Isabelle commands are written in bold font. Finally we describe our work in progress on the proof of soundness in Isabelle and relate to other works on formalization.

## 2 Normalized Formulas and Axiomatic System

The formulation for the axiomatic system by Monk [3] uses notions and expressions from Tarski [5], and Kalish and Montague [4]. In the context of our work, the notion of *normalized* formulas is the most important notion to understand. It relies on a logical validity in first-order predicate logic with equality that transforms arbitrary predicates into a logically equivalent formula where the predicate variables can be replaced with the arity of the predicate. For example, any binary predicate can be transformed as follows.

\[
P(x, y) \equiv \forall v_0 (v_0 = x \rightarrow \forall v_1 (v_1 = y \rightarrow P(v_0, v_1)))
\]

In the formula on the right-hand side, the variables \( v_0 \) and \( v_1 \) can be considered implicit arguments of \( P \) so that it is only necessary to represent \( P \) by its name and its arity.

To illustrate consider the case of stating that \( x + y = y + x \) where \( x \) and \( y \) are natural numbers. Let \( \text{Plus}(x, y, z) \) be true iff \( x + y = z \) where \( x, y \) and \( z \) are natural numbers. The statement can then be written as \( \forall x, y, z (\text{Plus}(x, y, z) \leftrightarrow \text{Plus}(y, x, z)) \). To get the normal form for Monk’s formulation, we transform it
into an equivalent expression as outlined above. Note that the term lists of the Plus predicates in the new formula are syntactically identical.

\[ \forall x, y, z (\text{Plus}(x, y, z) \leftrightarrow \text{Plus}(y, x, z)) \equiv \]
\[ \forall x, y, z (\forall v_0(v_0 = x \rightarrow \forall v_1(v_1 = y \rightarrow \forall v_2(v_2 = z \rightarrow \text{Plus}(v_0, v_1, v_2)))) \leftrightarrow \]
\[ \forall v_0(v_0 = y \rightarrow \forall v_1(v_1 = x \rightarrow \forall v_2(v_2 = z \rightarrow \text{Plus}(v_0, v_1, v_2)))) \]

Monk defines the axiomatic system \( A_1 \) shown in Table 1, where \( A, B, \) and \( C \) are normalized formulas and \( x, y, \) and \( z \) are variables. The occurrence check in (C5') for a predicate with arity \( n \) takes \( n \) comparisons due to the equalities in the transformation.

**Axioms**

(C1) \( (A \rightarrow B) \rightarrow (B \rightarrow C) \rightarrow A \rightarrow C \)

(C2) \( A \rightarrow \neg A \rightarrow B \)

(C3) \( (\neg A \rightarrow A) \rightarrow A \)

(C4) \( \forall x(A \rightarrow B) \rightarrow \forall xA \rightarrow \forall xB \)

(C5') \( A \rightarrow \forall xA \quad x \) does not occur in \( A \)

(C5') \( \neg \forall xA \rightarrow \forall x \neg \forall xA \)

(C5') \( \forall x \forall yA \rightarrow \forall y \forall xA \)

(C6) \( \neg \forall x(x = y) \)

(C7) \( x = y \rightarrow x = z \rightarrow y = z \)

(C8) \( x = y \rightarrow A \rightarrow \forall x(x = y \rightarrow A) \quad x \) is different from \( y \)

**Rules**

(R1) From \( A \) and \( A \rightarrow B \) infer \( B \)

(R2) From \( A \) infer \( \forall x A \)

**Table 1.** Monk’s axiomatic system \( A_1 \) for first-order predicate logic with equality. Note that (C5') only requires checking for occurrence, no matter if \( x \) is bound or free.

### 3 On Simplicity Compared to Natural Deduction

In this section we compare Monk’s formulation to natural deduction, as presented in a popular textbook on logic in computer science [11], in order to further clarify the notion of simplicity. To begin with, we highlight the use of substitution in natural deduction:

Given a variable \( x \), a term \( t \) and a formula \( \phi \) we define \( \phi[t/x] \) to be the formula obtained by replacing each free occurrence of variable \( x \) in \( \phi \) with \( t \).

On top of this definition there is a definition of what it means that ‘\( t \) must be free for \( x \) in \( \phi \):
Given a term $t$, a variable $x$ and a formula $\phi$, we say that $t$ is free for $x$ in $\phi$ if no free $x$ leaf in $\phi$ occurs in the scope of $\forall y$ or $\exists y$ for any variable $y$ occurring in $t$.

Having a definition of substitution the natural deduction rules are defined as follows:

\[
\begin{align*}
\begin{array}{c}
\neg \phi \\
\vdots \\
\hline \\
\phi \\
\end{array} & \quad \text{PBC} & \begin{array}{c}
\phi \\
\phi \rightarrow \psi \\
\hline \\
\psi \\
\end{array} & \quad \to E & \begin{array}{c}
\phi \\
\hline \\
\psi \\
\end{array} & \quad \to I \\
\end{align*}
\]

\[
\begin{align*}
\begin{array}{c}
\phi \\
\vdots \\
\hline \\
\chi \\
\end{array} & \quad \phi \lor \psi & \begin{array}{c}
\psi \\
\vdots \\
\hline \\
\chi \\
\end{array} & \quad \psi \lor \chi & \quad \lor E & \begin{array}{c}
\phi \\
\hline \\
\lor \psi \\
\end{array} & \quad \lor I_1 & \begin{array}{c}
\psi \\
\hline \\
\lor \psi \\
\end{array} & \quad \lor I_2 \\
\end{align*}
\]

\[
\begin{align*}
\begin{array}{c}
\phi \\
\hline \\
\phi \land \psi \\
\phi \land \psi \\
\hline \\
\phi \\
\end{array} & \quad \land E_1 & \begin{array}{c}
\psi \\
\hline \\
\phi \land \psi \\
\phi \land \psi \\
\hline \\
\psi \\
\end{array} & \quad \land E_2 & \begin{array}{c}
\phi \\
\hline \\
\phi \land \psi \\
\end{array} & \quad \land I \\
\end{align*}
\]

\[
\begin{align*}
\begin{array}{c}
\exists x \\
\vdots \\
\hline \\
\chi \\
\end{array} & \quad \exists x \phi & \begin{array}{c}
\chi \\
\vdots \\
\hline \\
\chi \\
\end{array} & \quad \exists E & \begin{array}{c}
\phi[t/x] \\
\hline \\
\exists x \phi \\
\end{array} & \quad \exists I \\
\end{align*}
\]

\[
\begin{align*}
\begin{array}{c}
\forall x \phi \\
\vdots \\
\hline \\
\chi \\
\end{array} & \quad \forall E & \begin{array}{c}
\chi \\
\vdots \\
\hline \\
\chi \\
\end{array} & \quad \forall E & \begin{array}{c}
\phi[t/x] \\
\hline \\
\forall x \phi \\
\end{array} & \quad \forall I \\
\end{align*}
\]

Side conditions to rules for quantifiers:

- $\exists E$: $x_0$ does not occur outside its box (and therefore not in $\chi$).
- $\exists I$: $t$ must be free for $x$ in $\phi$.
- $\forall E$: $t$ must be free for $x$ in $\phi$.
- $\forall I$: $x_0$ is a new variable which does not occur outside its box.

In addition there is a special copy rule:
A final rule is required in order to allow us to conclude a box with a
formula which has already appeared earlier in the proof. [...] The copy
rule entitles us to copy formulas that appeared before, unless they depend
on temporary assumptions whose box has already been closed.

For truth, negation and biimplication, the following equivalences can be used
where \( A \) and \( B \) are arbitrary formulas:

\[
\begin{align*}
\top & \equiv \bot \to \bot \\
\neg A & \equiv A \to \bot \\
A \leftrightarrow B & \equiv (A \to B) \land (B \to A)
\end{align*}
\]

There are a number of ways that the above formulation of natural deduction
can be compared to Monk’s formulation in terms of the simplicity of checking
correctness of a proof:

**Checking for free occurrence of variables** The rules for the quantifiers rely
on side conditions that rely on checking for free occurrence of variables in a
formula. In Monk’s formulation there are no side conditions that rely on this
check. There is a check for occurrence of variables but without the condition
that they must be free.

**Substitution** In addition to the above, the side condition for universal quanti-
tifier introduction also relies on substitution with a variable that does not
occur outside its box. Axiom scheme C5\(^1\) in Monk’s formulation relies on
checking for occurrence of a variable in a formula but it does not rely on
substitution.

**Copy rule** The copy rule does not have any immediately corresponding rule
in Monk’s formulation. Checking correctness of the application of the copy
rule involves checking if the formula in question has appeared earlier in the
proof outside a box that has already been closed. There is some resemblance
to applying a rule in Monk’s formulation in that it also involves referring to
some previously appearing formulas but there is no copying involved.

4 On Simplicity Compared to Sequent Calculus

Having looked at how Monk’s formulation compares to natural deduction in
terms of simplicity, we make a similar comparison with sequent calculus as pre-
sented by Tom Ridge [15].

Sequent calculus is a proof system in which the intention is to start from the
goal formula to be proven and then apply rules that break down the proof into
subgoals until there are no more goals left to be proven. Initially the goal formula
is transformed into negation normal form, yielding a formula in which the only
operators are \( \exists, \forall, \land, \lor \), or \( \neg \) and all negations are applied to predicates. Let
the initial current sequent be a singleton list containing this formula. A proof
is then a list of applications of the rules shown below to the current sequent,
where \( \Gamma \) and \( \Delta \) are possibly empty lists. Applying any of the two rules in the top
(marked with a *) removes a sequent and the other six rules replace the current sequent with the sequent(s) in the top. The proof is completed when there are no more sequents left.

The intention behind the system is that you start with the rules for \( \exists, \forall, \land, \) and \( \lor \) until the top four rules are applicable. The two rules \( \text{NoAx} \) and \( \overline{\text{NoAx}} \) are for “skipping” through formulas in the sequent either until one of the rules \( \text{Ax} \) or \( \overline{\text{Ax}} \) can remove the sequent or possibly never.

<table>
<thead>
<tr>
<th>Rule ((^* = \text{high priority}))</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \vdash P(v_1, \ldots, v_k), \Gamma, P(v_1, \ldots, v_k), \Delta ) ( \text{Ax}^* )</td>
<td>Leaf of the derivation tree.</td>
</tr>
<tr>
<td>( \vdash P(v_1, \ldots, v_k), \Gamma, P(v_1, \ldots, v_k), \Delta ) ( \overline{\text{Ax}}^* )</td>
<td>Leaf of the derivation tree.</td>
</tr>
<tr>
<td>( \vdash \Gamma, P(v_1, \ldots, v_k) ) ( \text{NoAx} )</td>
<td></td>
</tr>
<tr>
<td>( \vdash \Gamma, P(v_1, \ldots, v_k) ) ( \overline{\text{NoAx}} )</td>
<td></td>
</tr>
<tr>
<td>( \vdash \Gamma, A, B ) ( \lor )</td>
<td>The only branching rule.</td>
</tr>
<tr>
<td>( \vdash \Gamma, A \lor B, \Gamma ) ( \land )</td>
<td></td>
</tr>
<tr>
<td>( \vdash \Gamma, [v_i/x]A, (\exists x.A)^{i+1} ) ( \exists )</td>
<td>Superscripts are only relevant for this rule, and allow ([v_i/x]A) to be instantiated for all ( i ). ( v_r ) is a fresh free variable, chosen as ( r = \max(S)+1 ), where ( S ) is the set of subscripts already used for the free variables in ( A ) (( r = 0 ) if there are no free variables in ( A )).</td>
</tr>
<tr>
<td>( \vdash \Gamma, [v_r/x]A ) ( \forall )</td>
<td></td>
</tr>
</tbody>
</table>

We note the following when comparing the above notion of sequent calculus and Monk’s axiomatic system in terms of simplicity of checking a proof:

**Checking for free occurrence of variables** Analogous to the comparison with natural deduction, the side condition for \( \forall \) rely on checking for free occurrence of a variable in formula \( A \) whereas Monk’s axiomatic system only relies on checking for occurrence.
Substitution Similarly, the rules for $\exists$ and $\forall$ rely on substitution whereas Monk’s axiomatic system does not.

Scope In sequent calculus all rules only refers to the current sequent. Specifically the two rules $\text{NoAx}$ and $\overline{\text{NoAx}}$ are used for changing the current sequent to the head of $I$. In comparison, the rules in Monk’s axiomatic system refer to any of the previously appearing formulas.

5 Syntax and Semantics

In order to define the axiomatic system in Isabelle, we first define the syntax of normalized formulas. We define the syntax as a type using the Isabelle keyword `datatype` with two `abbreviation`s for truth and falsity. Thus values of this type represent normalized formulas. Note that the `nat` in the constructor for a predicate refers to the arity of the predicate, while the `nats` in the constructors for equality and universal formulas refer to variables.

```
datatype form = Pre string nat | Eq nat nat | Neg form | Imp form form | Uni nat form
```

```
abbreviation (input) Falsity ≡ Uni 0 (Neg (Eq 0 0))
abbreviation (input) Truth ≡ Neg Falsity
```

Although only the syntax is necessary in order to define the axiomatic system, we define the semantics next in order to show the meaning of the normalized formulas. The semantics is defined as a recursive function: it takes an environment, an interpretation, and a formula as input and calculates a truth value. The cases for equality and complex formulas are straightforward but the case for predicates follows from the normalized form. In this form, the number denotes the arity $n$ of the predicate and thus implies that the variables for the predicate are $v_0, ..., v_{n-1}$. Thus we can evaluate a predicate by mapping $v_0, ..., v_{n-1}$ to terms with the environment and then map the predicate name and terms to a truth value with the interpretation. An auxiliary recursive function for calculating the list of variables is used in this mapping.

```
fun vars :: nat ⇒ nat list
where
  vars 0 = [] |
  vars n = (vars (n-1)) ++ [n-1]
```

```
fun semantics :: (nat ⇒ 'a) ⇒ (string ⇒ 'a list ⇒ bool) ⇒ form ⇒ bool
where
  semantics e g (Eq x y) ⟷ (e x) = (e y) |
  semantics e g (Pre p arity) ⟷ g p (map e (vars arity)) |
  semantics e g (Neg f) ⟷ ¬ semantics e g f |
  semantics e g (Imp p q) ⟷ (semantics e g p → semantics e g q) |
  semantics e g (Uni x p) ⟷ (∀ t. (semantics (e(x:=t)) g p))
```
6 Axioms and Rules

Having definitions for syntax and semantics of normalized formulas, we define
the axiomatic system and proof system.

To begin with, we define a recursive function for the occurrence check. The
function should return true only if the given variable does not occur in the
given formula. Recall that predicate formulas are defined only in terms of arity
as described earlier. Thus in the occurrence check for predicate $Pre p n$ with
variable $x$, it suffices to check if $x < n$. In order to save space the body of the
function has been omitted.

\[
\text{fun not-occurs-in :: nat ⇒ form ⇒ bool}
\]

Next we define a type that represents theorems. The idea is that the axiomatic
system produces values of this type from its axiom schemes and rules, and thus
the axiomatic system produces theorems. Inapplicable use of the rules and axiom
schemes with side conditions will result in trivial truth.

\[
\begin{align*}
\text{datatype thm} &= \text{Thm(concl: form)} \\
\text{abbreviation (input) fail-thm} &= \text{Thm Truth}
\end{align*}
\]

The two rules of the axiomatic system are defined as functions. Given the
input theorems they produce a new theorem. We define the rules using the
Isabelle keyword \texttt{definition}.

\[
\begin{align*}
\text{definition modusponens :: thm ⇒ thm ⇒ thm} \\
&\text{where} \\
\text{modusponens s s' ≡ case concl s of Imp p q ⇒} \\
&\text{let p' = concl s' in if p = p' then Thm q else fail-thm} \\
&\text{| - ⇒ fail-thm}
\end{align*}
\]

\[
\begin{align*}
\text{definition gen :: nat ⇒ thm ⇒ thm} \\
&\text{where} \\
\text{gen x a ≡ Thm (Uni x (concl a))}
\end{align*}
\]

We define the ten axioms schemes in a similar manner. These are functions
that produce a theorem given a number of formulas. For the axiom schemes
without side conditions it suffices that the formulas are well-founded (being of
the type \texttt{form}). For example for C1 and C2 we have

\[
\begin{align*}
\text{definition c1 :: form ⇒ form ⇒ form ⇒ thm} \\
&\text{where} \\
c1 p q r ≡ \text{Thm (Imp (Imp p q) (Imp (Imp q r) (Imp p r)))}
\end{align*}
\]

\[
\begin{align*}
\text{definition c2 :: form ⇒ thm} \\
&\text{where} \\
c2 p ≡ \text{Thm (Imp (Imp (Neg p) p) p)}
\end{align*}
\]

For the axiom schemes with a side condition we additionally require that
the formulas satisfy the side condition. This is done by using if-statements with
fail-thm as the failure value. For (C5\textsuperscript{2}) we perform the occurrence check with \textit{not-occurs-in} described earlier, and for (C8) we use non-equality. For example for C8 we have

\begin{verbatim}
definition c8 :: nat ⇒ nat ⇒ form ⇒ thm where
  c8 x y p ≡ if x ≠ y then Thm (Imp (Eq x y) (Imp p (Imp (Eq x y) p))) else fail-thm
\end{verbatim}

Thus far we have defined the axioms schemes and rules of the axiomatic system. We tie them all together in a proof system that given a formula returns true if the axiomatic system can produce a theorem with it and false otherwise. In the following section we define soundness of the axiomatic system in terms of this proof system. In Isabelle, we define the proof system inductively with the \textbf{inductive}-keyword. The annotation (\(\vdash \cdot \theta\)) at the end states that \(\vdash\) can be used as notation for OK with precedence 0 (lowest). The inductive definition states the theorems that follow from the axiomatic system. The first two cases state that the formulas of theorems produced by the rules follow from the axiomatic system. The remaining ten cases state that the formulas produced by the axiom schemes follow from the axiomatic system.

\begin{verbatim}
inductive OK :: form ⇒ bool (\vdash \cdot \theta)
where
  case-modusponens:
    \(\vdash \text{concl } f \implies \vdash \text{concl } f' \implies \vdash \text{concl (modusponens } f f')\) |
  case-gen:
    \(\vdash \text{concl } f \implies \vdash \text{concl (gen - } f)\) |
  case-c1:
    \(\vdash \text{concl (c1 - -)\) |}
  case-c2:
    \(\vdash \text{concl (c2 -)\) |}
  case-c3:
    \(\vdash \text{concl (c3 - -)\) |}
  case-c4:
    \(\vdash \text{concl (c4 - -)\) |}
  case-c5-1:
    \(\vdash \text{concl (c5-1 -)\) |}
  case-c5-2:
    \(\vdash \text{concl (c5-2 -)\) |}
  case-c5-3:
    \(\vdash \text{concl (c5-3 - -)\) |}
  case-c6:
    \(\vdash \text{concl (c6 - -)\) |}
  case-c7:
    \(\vdash \text{concl (c7 - -)\) |}
  case-c8:
    \(\vdash \text{concl (c8 - -)\) |}
\end{verbatim}

Thus we have defined an axiomatic system for first-order predicate logic with normalized formulas in Isabelle based on Monk’s formulation in \[3\], and a proof
system that we can proceed to prove soundness of and thus certify soundness of the axiomatic system.

7 Proving Soundness

Having defined the axiomatic system and the proof system for it in Isabelle, we can certify soundness by using interactive proof assistance in Isabelle. Our approach to proving soundness is by providing auxiliary lemmas about the definitions so that Isabelle automatically can check that soundness follows from those lemmas and the definitions. Isabelle assists by keeping track of the proof as it is written and what subgoals remain to be proven. These subgoals can be used as templates for the auxiliary lemmas which in some cases can be proven just by applying Isabelle library lemmas. For example the following lemma is solved by applying the Isabelle library function \texttt{fun-upd-idem}. The hypothesis \( x \neq y \) is not needed in the lemma. It is rather a product of using the subgoal from Isabelle directly as a template for the lemma.

\begin{verbatim}
lemma update-identity: \( x \neq y \rightarrow e \ x = e \ y \rightarrow \text{semantics} \ e \ g \ p \rightarrow \text{semantics} \ (e(x := e \ y)) \ g \ p \)
  by (simp add: \texttt{fun-upd-idem})
\end{verbatim}

Finally we define soundness of the proof system as follows.

\begin{verbatim}
theorem soundness: \( \vdash p \rightarrow \text{semantics} \ e \ g \ p \)
\end{verbatim}

The proof itself is a work in progress.

8 Related Work

In this section we consider other works in the literature that investigate formalizing proof systems for first-order logic in proof assistants, and how they differ from the work in this paper. The work in this paper follow in line with the work of Villadsen, Jensen and Schlichtkrull [9], who used Isabelle to formalize and generate code for the kernel of the LCF-style prover for first-order logic with equality by Harrison [8]. Their work is now also part of the Archive of Formal Proofs [10]. The work in this paper differs from their work in that we have investigated a formalization of Monk’s original formulation from 1965 rather than Harrison’s which is based on Monk and Tarski. Other work on computer assisted formalizations based on Monk and Tarski include the work of Normal Megill [14] on using the Metamath language for archiving, verifying, and studying mathematical proofs.

Earlier we compared Monk’s formulation with natural deduction and sequent calculus which have already been formalized in Isabelle. Villadsen, Jensen and Schlichtkrull [18] used Isabelle to prove soundness of a proof system for natural deduction. They developed the browser-based software NaDeA where one could make natural deduction proofs that would then be checked by the proof system.
verified by Isabelle. For sequent calculus, Tom Ridge and James Margetson [16] made a formalization of sequent calculus in Isabelle for which they proved soundness and completeness that is now also part of the Archive of Formal Proofs [15]. The formalization was later used as a starting point by Villadsen, Schlichtkrull and From [19] for a prover made with code-generation with the aim of being used for teaching logic and verification to bachelor computer science students.

Looking at work with Isabelle formalizations of other proof systems we have Blanchette et. al. [12] who investigated how codatatypes can be used for proving soundness and completeness of different kinds of proof systems for first-order logic in Isabelle. In comparison, we have focused on proving soundness of Monk’s formulation, and we have not investigated the use of codatatypes in proving soundness, although it is certainly worth considering in future work. There is also the work of Schlichtkrull [17] about a formalization of the resolution calculus for first-order logic that includes soundness and completeness. Looking even further at verification of proof systems in other frameworks we have the work of Kumar, Arthan, Myreen and Owens [13] on formalization of higher-order logic in HOL4.

9 Conclusion

The work in progress described in this paper shows the usefulness of working with axiomatic systems in a proof assistant. We have defined syntax and semantics of the first-order logic formulas and the axiomatic system. The proof assistant checks for type-correctness of the definitions and provides assistance in making a soundness proof of the axiomatic system. The soundness proof itself is a work in progress. In this way, we can certify soundness of Monk’s axiomatic system in Isabelle.

There is room for improvement in the current formalization. Currently we use “truth as failure” when a rule is not applicable as a mechanism to ensure soundness. Using a dependently-typed logical framework could potentially make this trick redundant and also simplify some of the rules. It is also worth considering how the formalization could include other terms than just variables. Going further, the work can be extended to show additional useful properties of the axiomatic system, such as completeness.

Acknowledgements

This work is part of the Industrial PhD project Hospital Planning with Multi-Agent Goals between PDC A/S and Technical University of Denmark. We are grateful to Innovation Fund Denmark for funding and the governmental institute Region H, which manages the hospitals in the Danish capital region, for being a collaborator on the project. We would like to thank PDC A/S for providing feedback on the ideas described in this paper. We would also like to thank Jørgen Villadsen and Anders Schlichtkrull for comments on a draft. Finally we would like to thank the referees for in-depth comments on the paper that we think helped improving it substantially.
References

Lexicographical Update for Belief Comparison Relations

Kai Li *

Peking University, Beijing, China,
likaiedemon@gmail.com

Abstract. In this paper, we discuss soft information and lexicographical update for probabilistic models and belief comparison relations for propositions. We first introduce comparison models that agrees with weight models, and then show that lexicographical updates have the effect as some weight-changing procedure, without using numbers. We also compare the difference between such updates and the traditional way of lexicographical updates for world-ordering semantics.

Keywords: epistemic logic, dynamic epistemic logic, relative likelihood, probabilistic logic, ordering semantics

1 Introduction

A rational way to represent beliefs is using probabilistic models. A probabilistic model is a structure $(W, P, V)$ where $W$ is a set, probability measure $P : \wp(W) \rightarrow [0, 1]$ is a finitely additive, i.e., for all disjoint $X, Y \subseteq W$, $P(X) + P(Y) = P(X \cup Y)$; and $V$ is a valuation that assigns to each propositional variable a subset of $W$. [Seg71] and [Gär75] have shown that there is complete logic for probabilistic models whose language contains only one modal operator “≽”. [VE15] connects probabilistic models to conditional beliefs, and [HII+13] suggests that we can use probabilistic models and the modal operator “≽” for expressing belief comparison “at least as likely as”. [Sco64] showed that a finite structure with a binary relation between propositions and so called Scottness properties would agree with some probabilistic model. We will call such structures comparison models in this paper. The binary relation in such comparison models exactly corresponds to the modal operator “≽”, and the logic proposed in [Gär75] is also complete w.r.t. to the class of comparison models. These works give us liberties to represent rational beliefs and modal expression “at least as likely as” without probabilistic function, and hence without numbers in models.

* The author thanks the China Scholarship Council (CSC) for a visiting grant to CWI, and CWI for hospitality during academic year 2016/2017. The author would also like to thank Jan van Eijck, Malvin Gattinger and Fernando R. Velazquez Quesada who provided suggestions that greatly improve the paper, and the author is immensely grateful to 3 “anonymous” reviewers for their comments on an earlier version of the manuscript.
There are also dynamic aspects of probabilistic models. [VBGK09] has proved that public announcement update for probabilistic models corresponds to Bayesian update in probabilistic theory. However not all “updates” for probabilistic models are Bayesian updates. An example is the learning process for Bayesian networks (see for instance [KF09] Part.III), where usually there is a (probably incorrect) probabilistic model in advance, and then agents would gradually modify this model into a more correct one by “learning” through observed data. Sometimes those data are partially observed, and hence cannot be taken as hard as public announcements. Indeed not every information can be trusted unconditionally, a kind of so-called soft information is treated in [VB07], which is not entirely reliable in the way that it affects beliefs without affecting knowledge. Furthermore, soft information updates do not involve numbers, but only revise “plausibility relations” between situations. Apparently the previous learning process for probabilistic models involves changing of numbers, however what if we do not have numerical estimation of the situation in the first place, but only have some vague impression of likelihoods between propositions? In this case we may use comparison models, in which no number occurs. Consider the following example for update with soft information:

My neighbor is in his yard. Since it is a normal day, I believe he is doing some recreation activity rather than being at risk. My friend just walked into my room and told me there is smoke in my neighbor’s yard and I believe he is telling the truth. Then I tend to believe that my neighbor is barbecuing rather than his yard is on fire. But I would also believe it is more likely that his yard is on fire than there is no smoke at all.

Smoke in his yard is a soft information because it was what I have been told but didn’t acquire directly. Thus I cannot completely rule out the possibility that my friend is not telling the truth. After this update, among possible smoke in his yard situations I would still believe that it is more likely that my neighbor is doing recreation activity (i.e., barbecuing) than being at risk (in this case being on fire). Furthermore, among possible situations without smoke, I may still believe doing recreation activity is more likely than being at risk. However, since I now strongly believe my neighbor’s yard is smoking, I would think even the least likely case among all smoking situations, i.e., that my neighbor’s yard is on fire, is more likely than there is no smoking at all.

In this paper we only consider a simple kind of soft information which is both strong and weak. It is strong in the sense that soft information $P$ would provide agents enough credence to believe that every proposition consistent with $P$ are more likely than any of those inconsistent with $P$, but would not affect agents’ beliefs any further. It is weak in the sense that agents would always suspect that it might not be the case $P$, i.e., they still know that not $P$ is possible. Thus we assume the following two postulates for soft information update, after a soft information $P$:

1. old relation remains if both propositions either entail $P$ or entail $\neg P$,
2. any proposition entails $P$ is more likely than not $P$.  

38
These postulates are similar in some sense to lexicographical update \( \uparrow \) for world-ordering models given by [VB07]:

\[ \uparrow P \] is an instruction for replacing the current ordering relation \( \leq \) between worlds by the following: all \( P \)-worlds become better than all \( \neg P \)-worlds, and within those two zones, the old ordering remains.

Nevertheless, it seems hard to directly transfer the results from world-ordering models to comparison models, because comparison models are stronger than world-ordering models, and [HII*13] argued that “this cannot be done on the basis of intuitive entailments”. [GdJ12] discussed soft information with models containing both world-ordering relations and comparison relations, however, lexicographical updates are still only used for world-ordering relations.

In section 2 we introduce comparison models, weight models and a complete system for them. In section 3 we discuss lexicographical update for comparison models and weight models. We compare them with lexicographical update for world-ordering models in section 4. The last section is conclusion and future work.

## 2 Comparison Model and Probability

In this section we introduce comparison models and weight models. We illustrate the connection between comparison models and weight models, and prove the completeness result for the class of comparison models.

**Definition 1** (Comparison Models) Given a countable set of proposition Prop, a comparison model \( \mathcal{M} \) is a tuple \((W, \succeq, V)\) where:

- \( W \) is a non-empty finite set of worlds;
- \( \succeq \) is a binary relation on \( \wp(W) \) with the following (Scottness) conditions for all \( X, Y \subseteq W \):
  1. \( W \succeq \emptyset \), where \( X \succeq Y \) is the abbreviation for \( X \succeq Y \) but not \( Y \succeq X \);
  2. \( X \succeq \emptyset \);
  3. \( X \succeq Y \) or \( Y \succeq X \);
  4. if \( \iota(A_1) + \iota(A_2) + \cdots + \iota(A_n) = \iota(B_1) + \iota(B_2) + \cdots + \iota(B_n) \) and \( B_i \succeq A_i \) for each \( i \in \{1, \ldots, n-1\} \), then \( A_n \succeq B_n \), where \( \iota : \wp(W) \to \{0, 1\}^W \) is the characteristic function such that for all \( A \subseteq W \) and \( w \in W \), \( \iota(A)(w) = 1 \) iff \( w \in A \).
- \( V : \text{Prop} \to \wp(W) \) is a valuation.

Because of totality, i.e., condition (3), \( X \succeq Y \) is equivalent to not \( Y \succeq X \).

A different kind of model has been used for probabilistic model. Instead of probabilistic functions, weight functions in finite models can also represent probabilistic measurements, as shown in [DR15] and [VE15].
Definition 2 (Weight Models and Agreement) A weight model $\mathcal{M}$ is a tuple $(W, L, V)$ where $W$ is a finite set of worlds, $L$ assigns to every $w \in W$ to a positive rational, and $V$ assigns to every $w \in W$ a subset of $P$. For each $X \subseteq W$, we use $\mathbb{L}(X) = \sum_{w \in X} L(w)$. A comparison relation $\succeq$ on $W$ agrees with $L$ if for all $X, Y \subseteq W$, $X \succeq Y$ iff $\mathbb{L}(X) \geq \mathbb{L}(Y)$. A comparison model $\mathcal{M} = (W, \succeq, V)$ agrees with a weight model $\mathcal{M}' = (W', L, V')$ if $W = W'$, $V = V'$ and $\succeq$ agrees with $L$.

Clearly each weight model is a probabilistic model. From each weight model $(W, L, V)$ we can also derive a probabilistic model $(W', P, V')$ by letting $W' = W$, $V' = V$ and for each $X \subseteq W$, $P(X) = \mathbb{L}(X) / \mathbb{L}(W)$. It is easy to check that $P$ is a probability measure, provided that $W$ is finite.

Theorem 3 (Scott Theorem) Each comparison model agrees with some weight model, and vice versa.

Proof. The proof was first given by Dana Scott, see Theorem 4.1 in [Sco64].

Definition 4 (Language and Semantics) Given a countable set of proposition Prop, the comparison language $\mathcal{L}_{\succeq}$ is defined by the following BNF:

$$\phi ::= \top | p | \neg \phi | (\phi \land \psi) | (\phi \triangleright \psi)$$

We use $\phi \triangleleft \psi$ for $\psi \triangleright \phi$ and $\phi \triangleright \psi$ for $\neg(\phi \triangleleft \psi)$. The universal modality $U$ is defined by $U\phi ::= \phi \triangleright \top$ and the existential modality $E\phi$ (dual of $U$) can be defined as $\phi \triangleright \bot$. The truth condition for $\triangleright$ in comparison models is given by:

$$\mathcal{M}, w \models \phi \triangleright \psi \iff [\phi]_M \succeq [\psi]_M$$

where $[\phi]_M = \{ w \in W \mid \mathcal{M}, w \models \phi \}$.

Note that if $\mathcal{M}$ is a weight model, then the truth condition for $\triangleright$ is

$$\mathcal{M}, w \models \phi \triangleright \psi \iff \mathbb{L}([\phi]_M) \geq \mathbb{L}([\psi]_M),$$

which intuitively says $\phi$ is more probable (or likely) than $\psi$. Thus $U\phi$ means that $\mathbb{L}([\phi]_M) \geq \mathbb{L}$, and then by our definition for weight function, $[\phi]_M = W$, i.e., $\phi$ is valid on $\mathcal{M}$. Therefore because of Theorem 3, $U$ is the universal modality for comparison models.

Proposition 1. Let $\mathcal{M} = (W, \succeq, V)$ be a comparison model, let $\mathcal{M}' = (W, L, V)$ be a weight model that $\mathcal{M}$ agrees with, let $\phi$ be a formula and let $w \in W$. Then $\mathcal{M}, w \models \phi$ iff $\mathcal{M}', w \models \phi$.

Proof. We prove by induction on the construction of $\phi$. The only crucial step is that $\mathcal{M}, w \models \phi \triangleright \psi$ iff $[\phi]_M \succeq [\psi]_M$ iff by Definition 2, $\mathbb{L}([\phi]_M) \geq \mathbb{L}([\psi]_M)$ iff $\mathcal{M}', w \models \phi \triangleright \psi$. 

40
[Seg71] and [Gär75]) introduced a schematic abbreviation $E$ to express
\[ \iota([\phi_1]_M) + \cdots + \iota([\phi_n]_M) = \iota([\psi_1]_M) + \cdots + \iota([\psi_n]_M). \]
Following [HM01] here we provide a different yet equivalent definition of $E$: for each $n \in \mathbb{N}$,
\[ \phi_1 \ldots \phi_n E \psi_1 \ldots \psi_n := \bigwedge_{k=1}^{n} U(\phi^{(k)} \leftrightarrow \psi^{(k)}) \]
where for all $k \in [1, \ldots, n]$ and $\chi \in \{\phi, \psi\},$
\[ \chi^{(k)} := \bigvee_{1 \leq i_1 < \cdots < i_k \leq n} (\chi_{i_1} \land \cdots \land \chi_{i_k}). \]

CL (comparison logic) calculus is the following system:

Taut  All instances of propositional tautologies
K      $U(\phi \to \psi) \to U\phi \to U\psi$
Ex     $(U(\phi \leftrightarrow \phi') \land U(\psi \leftrightarrow \psi')) \to (((\phi \geq \psi) \leftrightarrow (\phi' \geq \psi'))) $
Bot    $\phi \not\geq \bot$
BT     $\top \not> \bot$
Scott $\phi_1 \ldots \phi_n E \psi_1 \ldots \psi_n \to (\bigwedge_{i<n} (\phi_i \geq \psi_i)) \to (\psi_n \geq \phi_n)$ for each $n \in \mathbb{N}$

Rules:
\[ \frac{\phi \to \psi \phi}{\psi} \quad (\text{MP}) \quad \frac{\phi}{U\phi} \quad (\text{Nec-K}) \]

**Theorem 5** CL is sound and complete w.r.t. the class of comparison models.

**Proof.** We can show that each consistent CL-formula $\phi$ is satisfied by a finite canonical model with worlds containing propositional variables only occurring in $\phi$ that satisfies $\phi$. Furthermore Bot, BT, Tot and Scott correspond to Conditions 1-4 in Definition 1 respectively, and hence each those canonical model is a comparison model.

A more devious proof can be given as follows: [Gär75] showed that CL is sound and complete for weight models, and then by Theorem 3 and Proposition 2, CL is sound and complete for comparison models.

### 3 Lexicographical Update for Comparison Models

In this section, we discuss the lexicographical update for comparison models. Furthermore, we show that lexicographical update is equivalent to some weight-changing function in weight models. We also provide reduction axioms for lexicographical update.

**Definition 6** The lexicographical language $L_\phi$ is defined by the following BNF:
\[ \phi ::= \top | p | \neg \phi | (\phi \land \phi) | (\phi \geq \phi) | [\top] \phi. \]
Now we give the formal definition of lexicographical update for comparison models. Recall that after a lexicographical update $\lhd \phi$ with a soft information $\phi$, we assume the following:

1. old relation remains if both propositions either entail $\phi$ or entail $\neg \phi$, 
2. any proposition entails $\phi$ is more likely than one does not $\phi$.

**Definition 7** Let $M = (W, \succeq, V)$ be a comparison model. Given a soft information $\phi$, the updated comparison model $M^{\phi} = (W^{\phi}, \succeq^{\phi}, V^{\phi})$ is given by

- $W^{\phi} = W$,
- $\succeq^{\phi}$ is defined as for all $X, Y \subseteq W$, $X \succ^{\phi} Y$ iff
  - either $X \cap [\phi]_M \neq \emptyset$ iff $Y \cap [\phi]_M \neq \emptyset$ and $X \succ Y$
  - or $X \cap [\phi]_M \neq \emptyset$ and $Y \cap [\phi]_M = \emptyset$.
- $V^{\phi} = V$.

Lexicographical update $\lhd \phi$ for comparison model is equivalent to the following weight changing function $L^{\phi}$.

**Proposition 8** Let $M = (W, \succeq, V)$ be a comparison model, let $N = (W, \sqsubseteq, V)$ be a weight model agreeing with $M$, let $\lhd \phi$ be a lexicographical update for $M$, and let

$$L^{\phi}(w) = \begin{cases} L(w) + n & \text{if } w \in [\phi]_M \\ L(w) & \text{otherwise} \end{cases}$$

for some $n$. Then $\succeq^{\phi}$ agrees with $L^{\phi}$ iff $\min\{L(u) \mid u \in [\phi]_M\} + n > L([\neg \phi]_M)$.

**Proof.** Suppose $\succeq^{\phi}$ agrees with $L^{\phi}$. Let $w \in [\phi]_M$ such that $L(w)$ is minimal in $\{L(u) \mid u \in [\phi]_M\}$. Then by the definition of $\succeq^{\phi}$, $\{w\} \succ^{\phi} [\neg \phi]_M$, which implies

$$L(w) + n = L^{\phi}(w) > L^{\phi}([\neg \phi]_M) = L([\neg \phi]_M).$$

Suppose $L(w) + n > L([\neg \phi]_M)$ where $L(w)$ is minimal in $\{L(u) \mid u \in [\phi]_M\}$. Consider all $X, Y \subseteq W$.

- If both $X \cap [\phi]_M \neq \emptyset$ and $Y \cap [\phi]_M \neq \emptyset$, then $L^{\phi}(X) > L^{\phi}(Y)$ iff
  $$L^{\phi}(X \cap [\phi]_M) + L^{\phi}(X \cap [\neg \phi]_M) > L^{\phi}(Y \cap [\phi]_M) + L^{\phi}(X \cap [\neg \phi]_M)$$
  if
  $$L(X \cap [\phi]_M) + n + L(X \cap [\neg \phi]_M) > L(Y \cap [\phi]_M) + n + L(X \cap [\neg \phi]_M)$$
  iff $L(X) > L(Y)$ iff, by $\sqsubseteq$ agreeing with $L$, $X \succ Y$ iff $X \succ^{\phi} Y$.
- If both $X \cap [\phi]_M = \emptyset$ and $Y \cap [\phi]_M = \emptyset$, then because $\sqsubseteq$ agrees with $L$, we have $L(X) = L^{\phi}(X) > L^{\phi}(Y) = L(Y)$ iff $X \succ Y$ iff $X \succ^{\phi} Y$.
- If only one of $X, Y$ intersects with $[\phi]_M$, say $X \cap [\phi]_M \neq \emptyset$ and $Y \cap [\phi]_M = \emptyset$, then $X \succ^{\phi} Y$, and there is a $u \in X \cap [\phi]_M$ such that
  $$L^{\phi}(X) \geq L^{\phi}(u) = L(u) + n \geq L(w) + n > L([\neg \phi]_M) \geq L(Y) = L^{\phi}(Y).$$
Definition 9 **Semantics for $\models_{\theta}$:** let $\mathcal{M} = (W, \succeq, V)$ be a comparison model and let $w \in W$, 
\[ \mathcal{M}, w \models [\upharpoonright \phi] \psi \iff \mathcal{M}^{\phi}, w \models \psi. \]

A complete comparison lexicographical update logic (CLL) consists of the logic for CL and the following reduction axioms:

1. $[\upharpoonright \phi] p \leftrightarrow p$
2. $[\upharpoonright \phi] \neg \psi \leftrightarrow \neg [\upharpoonright \phi] \psi$
3. $[\upharpoonright \phi] (\alpha \wedge \beta) \leftrightarrow [\upharpoonright \phi] \alpha \wedge [\upharpoonright \phi] \beta$
4. $[\upharpoonright \phi] (\alpha \triangleright \beta) \leftrightarrow ((E([\upharpoonright \phi] \alpha \wedge \phi) \leftrightarrow E([\upharpoonright \phi] \beta \wedge \phi)) \wedge ([\upharpoonright \phi] \alpha \triangleright [\upharpoonright \phi] \beta)) \vee (E([\upharpoonright \phi] \alpha \wedge \phi) \wedge \neg E([\upharpoonright \phi] \beta \wedge \phi)).$

Note that the last reduction axiom is for operator $\triangleright$ instead of our primitive operator $\succ$, for it corresponds to our definition for $\triangleright_{\phi}^{\phi}$, which uses $\succ$ instead of $\triangleright_{\phi}^{\phi}$. Such reduction axiom would cause no problem, for $\alpha \triangleright \beta$ is the abbreviation of $\neg (\alpha \triangleright_{\phi} \beta)$, and using the axiom for negation, we are able to derive the reduction case for $\triangleright$.

**Theorem 1.** **CLL is complete w.r.t. $\models_{\theta}$.**

**Proof.** We only consider the soundness of the last reduction axiom

\[ [\upharpoonright \phi] (\alpha \triangleright \beta) \leftrightarrow ((E([\upharpoonright \phi] \alpha \wedge \phi) \leftrightarrow E([\upharpoonright \phi] \beta \wedge \phi)) \wedge ([\upharpoonright \phi] \alpha \triangleright [\upharpoonright \phi] \beta)) \vee (E([\upharpoonright \phi] \alpha \wedge \phi) \wedge \neg E([\upharpoonright \phi] \beta \wedge \phi)). \]

Let $\mathcal{M} = (W, \succeq, V)$ be a comparison model, let $w \in W$ and let $X = [\alpha]_{\mathcal{M}^{\phi}} = [\upharpoonright \phi] \alpha \bigm|_{\mathcal{M}}$ and $Y = [\beta]_{\mathcal{M}^{\phi}} = [\upharpoonright \phi] \beta \bigm|_{\mathcal{M}}$.

We can verify that the following equivalences hold:

- $\mathcal{M}, w \models_{\theta} [\upharpoonright \phi] (\alpha \triangleright \beta)$ iff $X \triangleright_{\phi} Y$.
- that $X \cap [\phi]_{\mathcal{M}} \neq \emptyset$ iff $Y \cap [\phi]_{\mathcal{M}} \neq \emptyset$, and $X \succeq Y$ is equivalent to that $\mathcal{M}, w \models_{\theta} E([\upharpoonright \phi] \alpha \wedge \phi) \leftrightarrow E([\upharpoonright \phi] \beta \wedge \phi)$ and $\mathcal{M}, w \models_{\theta} [\upharpoonright \phi] \alpha \triangleright [\upharpoonright \phi] \beta$.
- that $X \cap [\phi]_{\mathcal{M}} \neq \emptyset$ and $Y \cap [\phi]_{\mathcal{M}} = \emptyset$ is necessary and sufficient for that $\mathcal{M}, w \models_{\theta} E([\upharpoonright \phi] \alpha \wedge \phi)$ and $\mathcal{M}, w \models_{\theta} \neg E([\upharpoonright \phi] \beta \wedge \phi)$.

Using these equivalences and Definition 7 we can derive that the last reduction axiom is sound for comparison models.

4 **Update for World-Ordering Semantics**

Lexicographical update was first introduced by [VB07] for world-ordering semantics. We can define comparison relations between propositions from world-ordering relations. In this section we discuss two kinds of such comparison relations: $l$-lifting relations and $m$-lifting relations; and examine the differences of lexicographical updates between comparison relations $\succeq$ and these two relations respectively.
Definition 10 A world-ordering model $\mathcal{M}$ is a tuple $(W, \succeq, V)$ where $W$ is a non-empty finite set of worlds, $\succeq$ is a preorder on $W$ (reflexive and transitive) and $V$ is a valuation. We use $w \succeq u$ for $w \succeq u$ but not $u \succeq w$. Note that because we do not have totality for $\succeq$, $w \succeq u$ is not equivalent to not $u \succeq w$.

Given a lexicographical update $\uparrow \phi$ for $\mathcal{M}$, the updated world-ordering model $\mathcal{M}^{\uparrow \phi} = (W^{\uparrow \phi}, \succeq^{\uparrow \phi}, V^{\uparrow \phi})$ is given by

- $W^{\uparrow \phi} = W$,
- $\succeq^{\uparrow \phi}$ is defined as for all $w, u \in W$, $w \succeq^{\uparrow \phi} u$ iff
  - either $w \in [\phi]_{\mathcal{M}}$ iff $u \in [\phi]_{\mathcal{M}}$ and $w \succeq u$
  - or $w \in [\phi]_{\mathcal{M}}$ and $u \notin [\phi]_{\mathcal{M}}$,
- $V^{\uparrow \phi} = V$.

[Kra91] introduced $l$-lifting relations as relations between propositions extended from relations between worlds by letting proposition $X$ “better” than $Y$ if every $Y$-world has a greater $X$-world in the preorder $\succeq$. However such extension have some unwanted properties. For instance, it validates the following disjunction

$((\phi \succ \psi) \land (\phi \succ \chi)) \rightarrow (\phi \succ (\psi \lor \chi)),$

which is invalid over the class of weight models, and hence comparison models.

Formally given a world-ordering model $\mathcal{M} = (W, \succeq, V)$, $l$-lifting relation $\succeq_l$ is defined as follows:

$$X \succeq_l Y \iff \forall w \in Y \exists u \in X : u \succeq_l w.$$  

Proposition 11 Let $\mathcal{M} = (W, \succeq, V)$ be a world-ordering model, and let $\uparrow \phi$ be a lexicographical update for $\mathcal{M}$. Then $X \succeq_l^{\uparrow \phi} Y$ iff

- either $X \cap [\phi]_{\mathcal{M}} \neq \emptyset$, $Y \cap [\phi]_{\mathcal{M}} \neq \emptyset$ and $X \cap [\phi]_{\mathcal{M}} \succeq_l Y \cap [\phi]_{\mathcal{M}}$,
- or $X \cap [\phi]_{\mathcal{M}} = \emptyset$, $Y \cap [\phi]_{\mathcal{M}} = \emptyset$ and $X \succeq_l Y$,
- or $X \cap [\phi]_{\mathcal{M}} \neq \emptyset$ and $Y \cap [\phi]_{\mathcal{M}} = \emptyset$.

Proof. From left to right. Suppose $X \succeq_l^{\uparrow \phi} Y$.

- If both $X \cap [\phi]_{\mathcal{M}} \neq \emptyset$ and $Y \cap [\phi]_{\mathcal{M}} \neq \emptyset$, then for all $w \in X \cap [\phi]_{\mathcal{M}} \neq \emptyset$ and $u \in Y \cap [\phi]_{\mathcal{M}} \neq \emptyset$, the relation between $w$ and $u$ remains the same after update $\uparrow \phi$, and $w \succeq^{\uparrow \phi} u$ for each $u \in Y - [\phi]_{\mathcal{M}}$. It follows that $X \cap [\phi]_{\mathcal{M}} \succeq_l Y \cap [\phi]_{\mathcal{M}}$.
- If both $X \cap [\phi]_{\mathcal{M}} = \emptyset$ and $Y \cap [\phi]_{\mathcal{M}} = \emptyset$, then the relation between $X$-worlds and $Y$-worlds remains the same after update $\uparrow \phi$, which implies that $X \succeq_l Y$.
- The case that $X \cap [\phi]_{\mathcal{M}} \neq \emptyset$ and $Y \cap [\phi]_{\mathcal{M}} = \emptyset$ is just what we want.
- Last suppose that $X \cap [\phi]_{\mathcal{M}} = \emptyset$ and $Y \cap [\phi]_{\mathcal{M}} \neq \emptyset$. But this is impossible, for there is a $w \in Y \cap [\phi]_{\mathcal{M}}$ such that $w \succeq^{\uparrow \phi} u$ for each $u \in X$, contrary to $X \succeq_l^{\uparrow \phi} Y$.

From right to left.
First suppose $X \cap \downmodels[\phi]_M \neq \emptyset$, $Y \cap \downmodels[\phi]_M \neq \emptyset$ and $X \cap \downmodels[\phi]_M \models_l Y \cap \downmodels[\phi]_M$. Consider any $w \in Y$. If $w \notin \downmodels[\phi]_M$, then for each $u \in X \cap \downmodels[\phi]_M$, $u \models_l \phi$. If $w \in \downmodels[\phi]_M$, then by $X \cap \downmodels[\phi]_M \models_l Y \cap \downmodels[\phi]_M$ there is a $u \in X \cap \downmodels[\phi]_M$ such that $u \models_l w$, which implies $u \models_l \phi$. It follows that $X \models_l \phi$.

Second suppose $X \cap \downmodels[\phi]_M = \emptyset$, $Y \cap \downmodels[\phi]_M = \emptyset$ and $X \models_l Y$. Then since the relation between $X$-worlds and $Y$-worlds remains the same after update $\uparrow \phi$, we have $X \models_l \phi$.

Last suppose $X \cap \downmodels[\phi]_M \neq \emptyset$ and $Y \cap \downmodels[\phi]_M = \emptyset$. Because there is a $w \in X \cap \downmodels[\phi]_M$ such that $w \models_l \phi$ for all $u \in Y \subseteq \downmodels[\neg \phi]_M$, we can also obtain that $X \models_l \phi$.

As we can observe, $\models_l \phi$ is quite similar to $\models_l \neg \phi$, the only difference is that when $X \cap \downmodels[\phi]_M \neq \emptyset$ and $Y \cap \downmodels[\phi]_M \neq \emptyset$, for at least one $X$-world $x$, the truth condition $\models_l \phi$ is not $\models_l \neg \phi$. This is because after updating $\uparrow \phi$, $\models_l \phi$, $Y \cap \downmodels[\phi]_M$ is strictly “better” than $\models_l \neg \phi$, $Y \cap \downmodels[\neg \phi]_M$. Hence $X \cap \downmodels[\phi]_M \models_l Y \cap \downmodels[\phi]_M$ implies that for all $w \in Y \cap \downmodels[\neg \phi]_M$, $w' \models_l \phi$ for each $w' \in Y \cap \downmodels[\phi]_M$, and in turn $w \models_l \phi$ for some $u \in X \cap \downmodels[\phi]_M$.

Given a world-ordering model $\mathcal{M} = (W, \models_l, V)$ and $w \in W$, the truth conditions for $\models_l$ and $\models_l \phi$ are given by:

\[
\mathcal{M}, w \models \phi \models_l \psi \quad \text{iff} \quad \models[\phi]_\mathcal{M} \models_l \models[\psi]_\mathcal{M} \\
\mathcal{M}, w \models [\uparrow \phi] \models_l w \models \psi.
\]

A complete world-ordering lexicographical update logic (LL) consists of a complete logic for world-ordering models with l-lifting (see Theorem 7.5.1a [Hal03]):

K  \quad U(\phi \rightarrow \psi) \rightarrow U\phi \rightarrow U\psi  \\
BT \quad \neg(\bot \models_l \top)  \\
Tran \quad (\phi \models_l \psi) \rightarrow ((\psi \models_l \chi) \rightarrow (\phi \models_l \chi))  \\
J  \quad (\phi \models_l \psi) \land (\psi \models_l \chi) \rightarrow (\phi \models_l (\psi \lor \chi))  \\
Mon \quad U(\phi \rightarrow \psi) \rightarrow (\phi \models_l \psi)  \\
E  \quad E\phi \leftrightarrow \neg(\bot \models_l \phi)  \\

and the following reduction axioms:

1. $[\uparrow \phi] p \leftrightarrow p$
2. $[\uparrow \phi] \neg \psi \leftrightarrow \neg [\uparrow \phi] \psi$
3. $[\uparrow \phi] (\alpha \land \beta) \leftrightarrow [\uparrow \phi] \alpha \land [\uparrow \phi] \beta$

plus the following reduction axiom for $\models_l$:

\[
[\uparrow \phi] (\alpha \models_l \beta) \leftrightarrow (E([\uparrow \phi] \alpha \land \psi) \land E([\uparrow \phi] \beta \land \phi) \lor (E([\uparrow \phi] \alpha \land \phi) \land E([\uparrow \phi] \beta \land \phi) \land (E([\uparrow \phi] \models_l \beta \land \phi)) \lor \neg E([\uparrow \phi] \alpha \land \phi) \land \neg E([\uparrow \phi] \beta \land \phi)))
\]

**Theorem 2.** LL is complete for world-ordering models with l-lifting and world-ordering lexicographical update.
Proof. We only show that the reduction axiom for $\succsim$ is sound. Let $\mathcal{M} = (W, \succsim, V)$ be a world-ordering model, let $w \in W$ and let $X = [\alpha]_{\mathcal{M}^t \circ} = [[[\uparrow \phi] \alpha]_{\mathcal{M}}$, and $Y = [\beta]_{\mathcal{M}^t \circ} = [[[\uparrow \phi] \beta]_{\mathcal{M}}$. We can verify the following equivalences:

- $\mathcal{M}, w \models [\uparrow \phi] (\alpha \succsim \beta)$ iff $X \succsim_{\phi} Y$.
- $X \cap [\phi]_{\mathcal{M}} \neq \emptyset$, $Y \cap [\phi]_{\mathcal{M}} \neq \emptyset$ and $X \cap [\phi]_{\mathcal{M}} \succsim Y \cap [\phi]_{\mathcal{M}}$ iff $\mathcal{M}, w \models E([\uparrow \phi] \alpha \land \phi) \land E([\uparrow \phi] \beta \land \phi) \land ([\uparrow \phi] \alpha \supseteq [\uparrow \phi] \beta \land \phi)$.
- $X \cap [\phi]_{\mathcal{M}} = \emptyset$, $Y \cap [\phi]_{\mathcal{M}} = \emptyset$ and $X \succsim Y$ iff $\mathcal{M}, w \models \neg E([\uparrow \phi] \alpha \land \phi) \land \neg E([\uparrow \phi] \beta \land \phi) \land ([\uparrow \phi] \alpha \supseteq [\uparrow \phi] \beta)$.
- $X \cap [\phi]_{\mathcal{M}} \neq \emptyset$ and $Y \cap [\phi]_{\mathcal{M}} = \emptyset$ iff $\mathcal{M}, w \models E([\uparrow \phi] \alpha \land \phi) \land \neg E([\uparrow \phi] \beta \land \phi)$.

It follows Proposition 11 that the reduction axiom for $\succsim$ is sound world-ordering models with $l$-lifting and world-ordering lexicographical update.

To avoid unwanted properties of $l$-lifting relations, [HII+13] propose the following $m$-lifting relations. We can check that disjunction principle is invalid over $m$-lifting relations.

$m$-lifting relation $\succsim_m$ is defined as:

$$X \succsim_m Y \text{ iff there is an injection } f : Y \to X \text{ such that } \forall w \in Y: f(w) \succsim_m w.$$  

Note that an injection $f$ is a function such that $f(x) = f(y)$ only if $x = y$. As [HII+13] pointed out, $\succsim_m$ may allow incomparability between propositions. However, unlike $l$-lifting, it is hard to give the reduction axiom for $\succsim$ w.r.t. lexicographical update for $m$-lifting, if there is any. This is because $m$-lifting requires matching between propositions (by injection in definition), and after lexicographical update with some soft information $\phi$, there is no clear relation between old matching and possible new matching, especially when one proposition consists with $\phi$, and the other is in-consist with $\phi$. It seems that $m$-lifting is not flexible enough for soft information update.

5 Conclusion and Future Work

In this paper, we discussed soft information and lexicographical update for comparison models. Then we compared such updates with lexicographical updates for world-ordering models.

- We showed that lexicographical updates for comparison models are equivalent to some weight-changing functions for weight models. Thus lexicographical updates is a way of change agent’s probabilistic beliefs while without consulting numbers. Is there any other way to agent’s probabilistic beliefs without mentioning numbers?
- We presented a possible reduction axiom for lexicographical updates w.r.t. $l$-lifting relations. What about other models containing relations between propositions?
- Comparison models and weight models also have close relation to epistemic neighborhood models (see [VE15]). Are there any similar updates for such models? Furthermore, we only considered single agent case, what about multi-agent case?
References


Dialogue Games for Minimal Logic*

Alexandra Pavlova
Université Paris I Panthéon-Sorbonne
pavlova.alex22@gmail.com

Abstract. In the paper we define a dialogue system for Johansson’s minimal logic. We explore the connection between minimal dialogues and sequent calculus. A suitable sequent calculus for minimal logic is introduced. Then we describe an algorithm that transforms a winning strategy for Proponent in $D_{\min}$ into a sequent calculus for minimal logic.

Introduction

In the present paper we discuss some issues related to dialogue games and Johansson’s minimal logic. By dialogue games we understand dialogue logic of P. Lorenzen and K. Lorenz defining validity. Johansson’s minimal logic is quite well studied in the literature; however, we propose its dialogue characterisation and prove the relevant correspondence result. Major work has been done to prove correspondence between dialogue games and sequent calculi or natural deduction. Several authors proposed their proofs for the intuitionistic dialogues and the corresponding intuitionistic validity, such as Fermüller [2], Felscher [1], Sørensen and Urzyczyn [12]. One should mention that Rahman, Clerbout and Keiff [10] published their result for Minimal logic represented by Fitch-style natural deduction and corresponding dialogue logic which they define differently from this paper. An elegant version of proof has been proposed for both intuitionistic and classical logic by Alama, Knoks and Uckelman [5].

We propose a new system of minimal dialogue logic that we call $D_{\min}$, as well as a system from minimal sequent calculus $G_{\min}^a$. We prove correspondence between the two systems. The motivation for this work has two aspects. The first one represents a purely technical interest in exploring possible dialogue characterisations of various logics. Given many different ways to define intuitionistic and classical dialogues and to prove the relevant correspondence result, we are interested in setting up a minimal dialogue system. Though there already exists a variant of minimal dialogue logic [10], we come up with a different specification independently. Moreover, we give a proof of the correspondence result with respect to sequent calculus, not a Fitch-style natural deduction (as in [10]).

⋆ My thanks are due to Elena Lisanyuk and Iouri Netchitailov for stimulating and encouraging conversations and to Alberto Naibo for his relevant comments and assistance in improving of the paper.

1 We mostly use the terminology specified by Krabbe, cf. [7] which is different from that of [10], for instance we do not make use of a dialogue ”history.”
spect to sequent calculus, not a Fitch-style natural deduction (as in [10]). This paper. An elegant version of proof has been proposed for both intuitionistic and introduction and corresponding dialogue logic which they define differently from this published their result for Minimal logic represented by Fitch-style natural de-

and Urzyczyn [12]. One should mention that Rahman, Clerbout and Keiff [10]

tion. Several authors proposed their proofs for the intuitionistic dialogues and the correspondence between dialogue games and sequent calculi or natural deduc-

tion. Moreover, we give a proof of the correspondence result with re-

between the two systems. The motivation for this work has two aspects. The first

history

We use capital greek letters to refer to sets of formulae.

2 We use Gothic letters to indicate meta-language variables.

3 This is a restricted version of a more general axiom \( \mathfrak{A}, \Gamma \rightarrow \Theta, \mathfrak{A} \) where \( \mathfrak{A} \) is atomic and \( \Theta = \emptyset \). This restriction is used both for minimal and intuitionistic calculus, but not for the classical one.

4 We use capital greek letters to refer to sets of formulae.
where, for all rules, the succedent should contain exactly one formula\(^6\).

Multiset \(\mathcal{I}\) contains all the subformulae of the derived formula that are asserted in the antecedents of the sequents that constitute the inference. This corresponds to the minimal rule for dialogue logic that we discuss in 2. We claim a sequent calculus for minimal logic to be an intuitionistic calculus without the right weakening \((WR)\) of the form:

\[
\frac{\mathcal{I}}{\mathcal{D}} \quad \text{WR}\emptyset
\]

where \(\mathcal{D}\) is an arbitrary formula. It is easy to see that this rule corresponds to the Gentzen \(NJ\) rule of the form: \(\vdash \mathcal{D}\) [3]. However, as we have chosen the sequent calculus without separate structural rules, we modify the rules by imposing above the restriction on the succedent. Furthermore, we define negation as \(\neg \mathcal{A} \overset{\text{def}}{=} \mathcal{A} \supset \bot\). Thus, we can see that our rules for negation are just a special case of rules for implication:

\[
\frac{\mathcal{A}, \Gamma \rightarrow \bot}{\Gamma \rightarrow \neg \mathcal{A}} \quad \text{def}
\]

We provide our proof of correspondence result in section 3 for the sequent calculus with branching in \(\neg\text{Ax}\), but as one can see it is applicable to the variant without branching.

2 Minimal Dialogue Logic \(D^{\text{min}}\)

Here we introduce a dialogue interpretation for minimal logic that we call \(D^{\text{min}}\). We base our system on the Intuitionistic Dialogue Logic as it is presented by Krabbe in [7]. Dialogue is a two-player game about some formula with the Proponent \((P)\) whose task is to defend the formula in question and the Opponent \((O)\) who is responsible for giving a counterexample to the formula, thus showing that it is not generally valid. Normally, there are two following levels of rules in dialogue logic:

I Logical rules define the possible types of attacks and defences for each type of formula, i.e. containing specific logical operators. These rules show us how formulae can be criticised and defended. Thus, logical rules define the meaning of logical operators;

II Structural rules define the general course of game and its organisation. These rules define the exact protocol of communication, i.e. when each player can move and what type of move he is allowed to make.

\(^6\) For intuitionistic logic it would be that \(\Theta\) contains at most one formula; the classical logic does not have any restrictions on \(\Theta\).
There exist two possible types of moves: attacks and defences. An attack is a move of a player Y against one of propositions of the player X that can be executed either in a form a request to make an assumption; or in affirming the contrary. A defence is a response to an attack. A defence is always performed in the form of affirming a relevant proposition. We assume that the language of $D^{\text{min}}$ ($\mathcal{L}_{D^{\text{min}}}$) is similar to the $\mathcal{L}_{\text{min}}$, but without meta-language sequent sign.

**Definition 4 (Logical Rules)** The system $D^{\text{min}}$ has the following logical rules:

<table>
<thead>
<tr>
<th>Connective</th>
<th>Attack</th>
<th>Defence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X -! - A \land B$</td>
<td>$Y -? - \land_L$</td>
<td>$X -! - A$</td>
</tr>
<tr>
<td>$X -! - A \lor B$</td>
<td>$Y -? - \lor$</td>
<td>$X -! - A$</td>
</tr>
<tr>
<td>$X -! - A \supset B$</td>
<td>$Y -! - A$</td>
<td>$X -! - B$</td>
</tr>
<tr>
<td>$X -! - \neg A$</td>
<td>$Y -! - A$</td>
<td>$X -! - \bot$</td>
</tr>
<tr>
<td>$X -! - \forall X(f)$</td>
<td>$Y -? - \forall f/n$</td>
<td>$X -! - A[n/f]$</td>
</tr>
<tr>
<td>$X -! - \exists X(f)$</td>
<td>$Y -? - \exists f$</td>
<td>$X -! - A[n/f]$</td>
</tr>
</tbody>
</table>

where $A$ and $B$ are metavariables, $X$ and $Y$ variables for player (with $P$ and $O$ being the precise roles) with $X \neq Y$, $!$ and $?$ are used to represent assertion and demand respectively. In case of conjunction and universal quantification the choice is made by the attacker, whereas in case of disjunction and existential quantification the choice is made by the defender.

**Definition 5** A dialogue is a sequence of attacks and defences that begins with a finite (possibly empty) multiset $\Pi$ of formulae that are initially granted by $O$ and a finite (nonempty) multiset $\Delta$ of formulae that are initially disputed by $O^7$.

**Definition 6 (Structural rules)** The system $D^{\text{min}}$ has the following structural rules:

- **Start** The first move of the dialogue is carried out by $O$ and consists of an attack on (the unique) initially disputed formula $A^{8}$.

- **Alternation** Moves strictly alternate between players $O$ and $P$.

- **Atom** $P$ can affirm atomic formulae, including $\bot$, iff they were previously stated by $O$.

- **D11** If it is $X$’s turn and there is more than one attack by $Y$ that $X$ has not yet defended, only the most recent one may be defended$^9$.

---

$^7$ It is possible to dispute over a number of sormulae, however, in our account we assume that there is only one initially disputed formula.

$^8$ We count as a zero step the one where $P$ proposes a formula for the dispute.

$^9$ This is a rule for minimal and intuitionistic logic only. In classical logic any attack can be defended.
**D12** Any attack may be defended at most once\(^{10}\).

**Attack-rule** \(O\) can attack one and the same of \(P\)'s formulae only once, whereas \(P\) can attack \(O\)'s formula several times.

**Minimal rule** For each attack players must provide a defence\(^{11}\).

It is the structural rules, that define the type of logic corresponding to the dialogue system. The system \(G_{3\text{min}}^3\) corresponds to the minimal logic. To obtain this result we have added an additional rule that we call **minimal rule** to the set of previously proposed structural rules for intuitionistic logic as it is analogous to *ex falso quo datum* and we also had to modify the logical rule for negation for in intuitionistic and classical logic there is no possible defence for negation. The classical logic has the least number of rules and is obtained by rejecting the **minimal rule**, \(D11\) and \(D12\).

**Definition 7 (Ending)** The game ends if and only if the player whose turn it is to move has no legal move to make.

**Definition 8 (Winning conditions)** The game ends with \(P\) winning iff it is \(O\)'s turn and she has no possible move left to make.

The game ends with \(O\) winning iff it is \(P\)'s turn and she has no possible move left to make.

Unlike \([5]\), we do not have to define new ending rules and winning conditions, so we use traditional ones, like in \([7]\).

Each round in the dialogue consists of a number, moves of players\(^{12}\) (an assertion of either a formula / one of the symbolic attacks), a stance (either attack or defend), and a reference (a natural number referring to the indices of previous moves of the game). A round consists of an attack of \(X\) and a defence of \(Y\), or just an attack. To define validity, we need the notion of winning strategy.

**Definition 9** A dialogue tree \(T\) for a dialogue sequent \(\Pi \rightarrow \mathfrak{A}\) is a rooted directed tree whose nodes are rounds in a dialogue game such that every branch of \(T\) is a dialogue with initially granted formulas \(\Pi\) and initially disputed formula \(\mathfrak{A}\).

**Definition 10 (Winning Strategy)** A finite dialogue tree \(T'\) (\(T' \subset T\), where \(T\) is a full dialogue tree for the a dialogue sequent) is a winning strategy \(\tau\) for \(X\) if and only if each branch of the tree \(T'\) (regardless of any moves of \(Y\)) ends with the move of \(X\), i.e. player \(Y\) has no possible move to make.

---

\(^{10}\) This rule is applicable only for minimal and intuitionistic calculus, but in classical one \(P\) can repeat his defences.

\(^{11}\) Including the attack against the negation \(\neg\), as, in this paper, we have introduced a rule for defence against this attack.

\(^{12}\) Each round should contain at least one move.
3 The Correspondence between \( G^\text{min}_3 \) and \( D^\text{min} \)

Here we argue that inference in \( G^\text{min}_3 \) calculus represents a winning strategy for proponent in dialogue logic.

**Theorem 1 (Minimal validity).** Let \( A \) be any formula of propositional logic. The following conditions are equivalent:

1. There is a winning strategy for Proponent in dialogue \( D(A) \);
2. There exists a \( G^\text{min}_3 \) derivation of the formula \( A \) (i.e., \( \Gamma \rightarrow A \), where \( \Gamma \) is empty).

Furthermore, there exists an algorithm turning Proponent’s winning strategy into the \( G^\text{min}_3 \) derivation and visa versa.

We reformulate our theorem so that it can deal with any derivation (not necessarily of a theorem), i.e. that where \( \Gamma \) is not necessarily empty, and the corresponding dialogue with hypotheses.

**Theorem 2 (Correspondence result).** Every winning strategy \( \tau \) for Proponent with respect to \( D(A, \Gamma) \) (i.e., for a dialogue with initially disputed formula \( A \), where the Opponent initially grants the formulae in the multiset \( \Gamma \)) can be transformed into a \( G^\text{min}_3 \) derivation of \( \Gamma \rightarrow A \) and visa versa.

**Proof.** We prove theorem 2 in two steps, by establishing two lemmata.

**Lemma 1.** Every \( P \)'s winning strategy \( \tau \) for \( D(A, \Gamma) \) can be transformed into a \( G^\text{min}_3 \) derivation of \( \Gamma \rightarrow A \).

Now let us show that there exists an algorithm that transforms any particular dialogue into a branch of a sequent calculus. We shall use our table representation of a dialogue to be able to make use of the notion of rounds. Each round represents a sequence so that we can build out sequence from the table where 0-round represents the formula \( \mathcal{A} \) to be deduced (i.e., \( \Gamma \rightarrow \mathcal{A} \)). As the Opponent cannot repeat his attacks on previously attacked formulae of the Proponent, we do not keep the formulae whose subformulae have already been asserted in the succedent of the sequence\(^{13}\). On the contrary, we keep formulae that were asserted by the Opponent because the Proponent has a right to repeat his attack on a formula of the Opponent\(^{14}\).

**Proof.** Consider an arbitrary winning strategy \( \tau \) for the proponent.

**Claim.** For every round of the dialogue \( D(A, \Gamma) \) there is a \( G^\text{min}_3 \) deduction of the sequent corresponding to the dialogue sequent at this round \( \Theta \rightarrow \mathcal{A}_1 \), which consists of \( \Theta = \Gamma \cup \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n\} \), where \( \mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n \) are subformulae of \( A \) asserted by \( O \) and \( A_i \) represents a subformula of \( A \) asserted by \( P \). Note that \( A_i \) might be empty.

\(^{13}\) This requirement corresponds to the restrictions imposed on the succedent in intuitionistic logic, where we cannot have more than one formula in the succedent.

\(^{14}\) This corresponds to the rules of \( G^\text{min}_3 \) where we keep all the formulae in the antecedent.
Proof. We prove this claim by induction on the depth \( d \) of \( \tau \).

The base case: \( d = 1 \). In this case the game terminates at the round \#1. So in this round Proponent moves last according to the winning conditions 8. Thus, the Proponent has asserted an atomic formula \( A \) because otherwise (if the formula asserted by the proponent at the round \#1 were not atomic) the Opponent would have been able to move by attacking a complex formula. As \( A \) is atomic, the Opponent must have already granted \( A \) (according to the atom rule 6) either with his move in the first round or in \( \Gamma \). Thus, our dialogue sequent at round \#1 is \( \forall, \Gamma \rightarrow A \) (or simply \( \Gamma \rightarrow A \) if \( A \in \Gamma \)).

\#1: \( A := p \supset p \) is an example of a formula whose dialogue ends in round:

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>( (0) ) ( p \supset p )</td>
</tr>
<tr>
<td>1</td>
<td>( (1) ) ( p ) ( \text{Att.0} )</td>
<td>( (2) ) ( p ) ( \text{Def.1} )</td>
</tr>
</tbody>
</table>

We can easily transform this into the \( G_{3a}^{\text{min}} \) deduction. According to our algorithm we associate a sequence to each round. In our example we have only two rounds, so we associate the sequence \( \Gamma \rightarrow A \) with the 0-round (in our example it is \( \rightarrow p \supset p \) which is associated with the 0-round where the Proponent states formula \( p \supset p \) with \( \Gamma = \emptyset \)) and the sequence \( A, \Gamma \rightarrow A \) to the first round (in our example it is \( p \rightarrow p \) which is associated with the first round and the statement \( p \) representing the attack in the form of assertion by the opponent and the defence \( p \) of the proponent). Thus, we get the following \( G_{3a}^{\text{min}} \) deduction:

\[
\frac{p \rightarrow_1 p}{\rightarrow_0 p \supset p} \hspace{1cm} (\supset^{S+})
\]

The induction hypothesis: we assume that the claim holds for all \( n \leq d \) and show that it also holds at \( d + 1 \) step.

The inductive step: we proceed by analysing the cases according to the form of the formula that is defended or attacked by \( P \).

1. \( P \) defends \( A \land B \). So our previous dialogue sequent has the following form: \( \Theta \rightarrow A \land B \). So if \( O \) can attack \( A \land B \) by either ? \(- \land_L \) or ? \(- \land_R \). So if \( O \) makes the move ? \(- \land_L \), then at some point \( P \) will have to assert \( A \) (according to the minimal rule), and if \( O \) makes the move ? \(- \land_R \), then \( P \) will have to assert \( B \).

   (a) If \( A \) (and \( B \) accordingly) is atomic, then our sequent has the form \( C, \Theta \rightarrow C \) (where \( C \) is atomic formula \( A := C \), or \( B := C \), and \( C \in \Theta \)), and we have reached our end as there should already be an assertion of \( C \) made by \( O \) in order for \( P \) to be able to assert \( C \) (cf. the base case). By induction, we get that all the previous steps \( d \) can be transformed into the sequent form, thus the whole dialogue can be transformed into \( G_{3a}^{\text{min}} \) deduction.

   (b) If \( A \) (and \( B \)) is not atomic, then we get a dialogue of the form \( \Theta \rightarrow A \) (\( \Theta \rightarrow B \)) and then we continue our derivation with the formula \( A \) (\( B \)). According to our inductive step there already exists a derivation of \( A \) (\( B \)), and thus our transition from round \( d \) to round \( d + 1 \) can be transformed into the \( \land^{S+} \) rule of \( G_{3a}^{\text{min}} \), which he choses himself.
2. \( P \) attacks \( \mathcal{A} \land \mathcal{B} \). Then the previous dialogue sequent has the form \( \Theta, \mathcal{A} \land \mathcal{B} \rightarrow \mathcal{D} \) (where \( \mathcal{D} \) might be empty). This case is analogous to the cases 1a and 1b.

3. \( P \) defends \( \mathcal{A} \lor \mathcal{B} \). Thus, the previous dialogue sequent has the following form: \( \Theta \rightarrow \mathcal{A} \lor \mathcal{B} \). \( O \) has only one possible attack \( \land \rightarrow \lor \). The proponent has two possible defences: \( \mathcal{A} \) or \( \mathcal{B} \). Thus, in the next step we get either \( \Theta \rightarrow \mathcal{A} \) or \( \Theta \rightarrow \mathcal{B} \), thus, this dialogue sequent can be transformed into the \( G_3^{\min} \) sequences:

\[
\begin{align*}
\Theta & \rightarrow \mathcal{A} \\
\Theta & \rightarrow \mathcal{A} \lor \mathcal{B} \\
\Theta & \rightarrow \mathcal{B}
\end{align*}
\]

Either way the game proceeds on the dialogue sequent \( \Theta \rightarrow \mathcal{A} \) or \( \Theta \rightarrow \mathcal{B} \) (and so does the \( G_3^{\min} \) inference). And according to our inductive step there exists a \( G_3^{\min} \) inference for the dialogue \( \mathcal{D}(\mathcal{A}, \Gamma) \) (with the strategy \( \tau \) of the length \( D \)), where \( A = \mathcal{A} \) or \( A = \mathcal{B} \). We have shown that this also holds for the dialogue \( \mathcal{D}(\mathcal{A} \lor \mathcal{B}, \Gamma) \) (for \( \tau \) of the length \( d + 1 \)).

4. \( P \) attacks \( \mathcal{A} \lor \mathcal{B} \). Then the previous dialogue sequent has the form \( \Theta, \mathcal{A} \lor \mathcal{B} \rightarrow \mathcal{D} \) (where \( \mathcal{D} \) might be empty). This case is analogous to the case 3.

5. \( P \) defends \( \mathcal{A} \supset \mathcal{B} \). At the previous round we get the dialogue sequent \( \Theta \rightarrow \mathcal{A} \supset \mathcal{B} \). As \( P \) defends the formula, this means that \( O \) has already attacked it. The only way to attack this formula is to state \( \mathcal{A} \). As we are in the minimal logic, we have a rule that all attacks (but for the one for the negation) should be defended (in other words, the rounds should be closed). And the only way to defend the formula in question against the attack is to assert \( \mathcal{B} \). Thus, at \( d \) we get to the dialogue sequent \( \mathcal{A}, \Theta \rightarrow \mathcal{B} \). We have proven our claim for \( d \), so by induction step we get that it should be true at \( d + 1 \) by showing the resulting inference:

\[
\begin{align*}
\mathcal{A}, \Theta & \rightarrow \mathcal{B} \\
\Theta & \rightarrow \mathcal{A} \supset \mathcal{B}
\end{align*}
\]

6. \( P \) attacks \( \mathcal{A} \supset \mathcal{B} \). This case is analogous to 5 but has some additional restrictions. We keep \( \mathcal{A} \supset \mathcal{B} \) as \( P \) can attack several times the formulae accepted by \( O \). We are tempted to write the following inference with formula \( \mathcal{D} \):

\[
\begin{align*}
\mathcal{B}, \mathcal{A} \supset \mathcal{B}, \Theta & \rightarrow \mathcal{A} \\
\mathcal{A} \supset \mathcal{B}, \Theta & \rightarrow \mathcal{D}
\end{align*}
\]

But then it is not minimal as the transformation implies \( WR \)-rule:

\[
\begin{align*}
\mathcal{B}, \mathcal{A} \supset \mathcal{B}, \Theta & \rightarrow \mathcal{A} \\
\mathcal{A} \supset \mathcal{B}, \Theta & \rightarrow \mathcal{D}
\end{align*}
\]

The attack of \( P \) on \( \mathcal{A} \supset \mathcal{B} \) should be transformed into the branching corresponding to the rule \( \supset^A \). Justification: an implication \( \mathcal{A} \supset \mathcal{B} \) says that given \( \mathcal{A} \) we can deduce \( \mathcal{B} \), so if it is stated by some player \( X \), in order for the player \( Y \) to challenge it, the player \( Y \) should state the antecedent \( \mathcal{A} \) and
the player $X$ should state then the consequent $B$. When $O$ states $A \supset B$, he claims the existence of a transition from $A$ to $B$. Hence, if we just wrote the sequence corresponding to the attack and defence as $B, \Gamma \rightarrow A$, we would claim that $A$ is the consequence of $B \cup \Gamma$, which, in case of $\Gamma = \emptyset$ is false. To clarify it we just say that $P$ affirms $A$ which might depend on some previous assertions of $O$ represented by $\Theta$ and the $O$ can defend it by stating $B$ that does not imply $A$ (i.e., as a separate branch), but which can imply a statement previously asserted by $P$. Thus we get the following branching:

$$
\begin{align*}
A \supset B, \Theta & \rightarrow A \quad \text{and} \quad B, A \supset B, \Theta \rightarrow A_i, \\
& \quad \text{and} \quad \text{A} \supset \text{B}, \Theta \leftarrow A_i.
\end{align*}
$$

7. $P$ defends $\neg A$. So at the previous round the dialogue sequent has the form $\Theta \rightarrow \neg A$ (step $d + 1$). the attack on the formula $\neg A$ represents an assertion of $A$. This case is analogous to the defence of an implication, as $\neg A \equiv A \supset \bot$ taking into account that $\bot$ is atomic.

8. $P$ attacks $\neg A$. Here we get the previous dialogue sequent $\neg A, \Theta \rightarrow \bot$. This case is analogous to 7 and 6, because it is a particular case of the $P$'s attack on implication.

Every round can be transformed into a sequent of a derivation. Thus, if we transform all dialogue winning strategies for the formula in question, we will get a $G_{\min}^{\text{ax}}$ deduction of the formula. The branching of $\land S^+$ and $\lor A^+$ in $G_{\min}^{\text{ax}}$ represents choices of the opponent that influence the (so they mark the branching of a strategy into two different ones) strategy of the proponent. If we transform our strategies as proposed in our claim, we shall see that each winning strategy end up with the $\text{axiom}$ of $G_{\min}^{\text{ax}}$. This is the case, because the $O$ has no moves to make if and only if all the formulæ asserted and not yet attacked by the $P$ are atomic (otherwise the $O$ could use the corresponding logical rule).

In order for $P$ to assert an atomic formula, it should be already stated by $O$, thus we get the $\text{axiom}$ of the from: $D, \Theta \rightarrow D$. Thus all winning strategies (for dialogues $\mathcal{D}(A, \Gamma)$) for a formula in question can be transformed into valid $G_{\min}^{\text{ax}}$ deductions.

**Lemma 2.** Every $G_{\min}^{\text{ax}}$ derivation of $\Gamma \rightarrow A$ can be transformed into a winning strategy $\tau_i$ for $\mathcal{D}(A, \Gamma)$.

**Proof.** To prove lemma 2 we show that each rule of $G_{\min}^{\text{ax}}$ can be transformed into a corresponding dialogue rule. Each $G_{\min}^{\text{ax}}$ corresponds to two rounds of the dialogue where round $\# m$ represents the initial formula and round $\# m + 1$ contains the attack (according to the logical rules) and defences (if there exist a possible defence according to the logical rules). To transform our $G_{\min}^{\text{ax}}$ rules into dialogue rules we shall map the antecedent of the sequent to the $O$ column and the consequent — to the $P$ column respectively. In the $O$ column we write only the formula that is being attacked only. All other formulæ in the antecedent of the sequent represent the previous assertions of the Opponent that can still be attacked by the Proponent. Let us show this correspondence for each of the rule independently:
1. Let us start with the rule \( \neg \vdash S^+ \):

\[
\neg A, \Gamma \rightarrow \bot \\
\Gamma \rightarrow \neg A \quad \neg \vdash S^+
\]

This can be mapped into the rule for attacking the negation (the attack is effectuated by \( O \), so \( \neg A \) is asserted by \( P \)). We associate the formula \( \neg A \) with the \( P \) column \( \# m \) and \( A \) with the \( O \) column \( \# m + 1 \). \( \Gamma \) represents \( O \)'s assertions made in rounds \( 0 \leq j \leq m \). So, we get:

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m - 1 )</td>
<td>( \cdots )</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( m )</td>
<td>( \Gamma )</td>
<td>( \neg A )</td>
</tr>
<tr>
<td>( m + 1 )</td>
<td>( A ) (Att.)</td>
<td>( \bot )</td>
</tr>
<tr>
<td>( m + 2 )</td>
<td>( \cdots )</td>
<td>( \cdots )</td>
</tr>
</tbody>
</table>

2. Consider the rule for negation introduction in the antecedent \( \neg \vdash A^+ \):

\[
\neg A, \Gamma \rightarrow A \quad \neg A, \Gamma, \bot \rightarrow \bot \\
\neg \vdash A, \Gamma \rightarrow \bot 
\]

Thus we get the following dialogue rule:

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m - 1 )</td>
<td>( \Gamma )</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( m )</td>
<td>( \neg A )</td>
<td>( A_i )</td>
</tr>
<tr>
<td>( m + 1 )</td>
<td>( \bot )</td>
<td>( A ) (Att.)</td>
</tr>
<tr>
<td>( m + 2 )</td>
<td>( \cdots )</td>
<td>( \cdots )</td>
</tr>
</tbody>
</table>

where \( A_i \) might be empty or represent an attack on the previous formula.

3. In the case of disjunction in the succedent \( \lor \vdash S^+ \) it is sufficient for one disjunct to satisfy the formula, and thus it is the proponent who chooses in the dialogue.

\[
\Gamma \rightarrow A \quad or \quad \Gamma \rightarrow B \\
\Gamma \rightarrow A \lor B \quad \lor \vdash S^+
\]

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m - 1 )</td>
<td>( \cdots )</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( m )</td>
<td>( \Gamma )</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( m + 1 )</td>
<td>( A \lor B )</td>
<td>( \cdots )</td>
</tr>
<tr>
<td>( m + 2 )</td>
<td>( \cdots )</td>
<td>( \cdots )</td>
</tr>
</tbody>
</table>

4. In the case of disjunction in the antecedent \( \lor \vdash A^+ \), both conjuncts should satisfy the formula (thus there is branching in \( G_{3}^{\min} \) derivation), thus, it is the opponent who chooses in the dialogue. Hence, the proponent should have a winning strategy for both disjuncts.

\[
A, A \lor B, \Gamma \rightarrow \Theta \\
B, A \lor B, \Gamma \rightarrow \Theta \\
\Gamma \rightarrow \Theta \\
\lor \vdash A^+
\]
5. The case for conjunction in the succedent $\wedge^{S+}$ is inverse with respect to the case of disjunction 3. Here both conjuncts should satisfy the formula, and thus, there is branching.

$$\Gamma \rightarrow \mathcal{A} \text{ and } \Gamma \rightarrow \mathcal{B} \quad \xrightarrow{\wedge^{S+}} \quad \Gamma \rightarrow \mathcal{A} \wedge \mathcal{B}$$

6. The case of conjunction in the antecedent $\wedge^{A+}$ is inverse to the disjunction 4. If something follows from at least one conjunct, then it follows from the conjunction.

$$\mathcal{A}, \mathcal{A} \wedge \mathcal{B}, \Gamma \rightarrow \Theta \quad \text{or} \quad \mathcal{B}, \mathcal{A} \wedge \mathcal{B}, \Gamma \rightarrow \Theta \quad \xrightarrow{\wedge^{A+}} \quad \mathcal{A} \wedge \mathcal{B}, \Gamma \rightarrow \Theta$$

7. The case of implication in succedent $\supset^{S+}$ is quite straightforward.

$$\mathcal{A}, \Gamma \rightarrow \mathcal{B} \quad \xrightarrow{\supset^{S+}} \quad \Gamma \rightarrow \mathcal{A} \supset \mathcal{B}$$

8. We have discussed the case of $\supset^{A+}$ proving claim 3. This case of branching in $G^{\supset}_{3} \cup a$ deduction does not correspond to the branching of strategies. Here both branches belong to one strategy and one dialogue. However, as we have no branching in our dialogue, we just associate two branches of this rule with one round. We associate the formula in the consequent with the $P$ column and the formulae in the antecedent with $O$ column:

$$\mathcal{A} \supset \mathcal{B}, \Gamma \rightarrow \mathcal{A} \text{ and } \mathcal{B}, \mathcal{A} \supset \mathcal{B}, \Gamma \rightarrow \Theta \quad \xrightarrow{\supset^{A+}} \quad \mathcal{A} \supset \mathcal{B}, \Gamma \rightarrow \Theta$$

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m-1$</td>
<td>$\Gamma$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$m$</td>
<td>$\mathcal{A} \lor \mathcal{B}$</td>
<td>$\mathcal{A}_1$</td>
</tr>
<tr>
<td>$m+1$</td>
<td>$? \lor \mathcal{B}$</td>
<td>$\mathcal{A}$</td>
</tr>
<tr>
<td>$m+2$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>

and

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m-1$</td>
<td>$\Gamma$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$m$</td>
<td>$\mathcal{A} \lor \mathcal{B}$</td>
<td>$\mathcal{A}_1$</td>
</tr>
<tr>
<td>$m+1$</td>
<td>$\mathcal{B} \lor \mathcal{A}$</td>
<td>$\mathcal{A}$</td>
</tr>
<tr>
<td>$m+2$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m-1$</td>
<td>$\Gamma$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$m$</td>
<td>$\mathcal{A} \lor \mathcal{B}$</td>
<td>$\mathcal{A}_1$</td>
</tr>
<tr>
<td>$m+1$</td>
<td>$\mathcal{B} \lor \mathcal{A}$</td>
<td>$\mathcal{A}$</td>
</tr>
<tr>
<td>$m+2$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m-1$</td>
<td>$\Gamma$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$m$</td>
<td>$\mathcal{A} \lor \mathcal{B}$</td>
<td>$\mathcal{A}_1$</td>
</tr>
<tr>
<td>$m+1$</td>
<td>$\mathcal{B} \lor \mathcal{A}$</td>
<td>$\mathcal{A}$</td>
</tr>
<tr>
<td>$m+2$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m-1$</td>
<td>$\Gamma$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$m$</td>
<td>$\mathcal{A} \lor \mathcal{B}$</td>
<td>$\mathcal{A}_1$</td>
</tr>
<tr>
<td>$m+1$</td>
<td>$\mathcal{B} \lor \mathcal{A}$</td>
<td>$\mathcal{A}$</td>
</tr>
<tr>
<td>$m+2$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Round</th>
<th>Opponent</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m-1$</td>
<td>$\Gamma$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$m$</td>
<td>$\mathcal{A} \lor \mathcal{B}$</td>
<td>$\mathcal{A}_1$</td>
</tr>
<tr>
<td>$m+1$</td>
<td>$\mathcal{B} \lor \mathcal{A}$</td>
<td>$\mathcal{A}$</td>
</tr>
<tr>
<td>$m+2$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>
6. The case of conjunction in the antecedent
\[ \land \]

5. The case for conjunction in the succedent
\[ \supset \]

7. The case of implication in succedent
\[ \supset \]

8. We have discussed the case of thus, there is branching.

If something follows from at least one conjunct, then it follows from the in
\[ \land \]

If something follows from at least one conjunct, then it follows from the in
\[ \land \]

\[ \text{Round} \quad \text{Opponent} \quad \text{Proponent} \]

\[ m - 1 \quad I \quad \cdots \]

\[ m \quad \mathcal{A} \supset \mathcal{B} \quad \Theta \]

\[ m + 1 \quad \mathcal{B} \quad \text{(Def.)} \quad \mathcal{A} \quad (\text{Att.}) \]

\[ m + 2 \quad \cdots \quad \cdots \]

With this rule we can face the problem of the order of steps after the branching. To identify the order we should make use of the rule forbidding \( P \) to assert atomic formulae that have not been yet asserted by \( O \).

The fact that for a valid derivation all rounds in the corresponding dialogue will be closed is guaranteed by the minimal rule of the dialogue (cf. 6). Furthermore, each valid derivation ends up with the axiom. This guarantees us that the dialogue has a winning strategy, because it ends with the atomic formula asserted by \( P \). It is so because \( O \) cannot attack atomic formulae (so there is nothing left for her to attack), and \( P \) could not assert it before \( O \)'s assertion of the corresponding atomic formula.

References


Resolution Calculi for Modal Logic and their Relative Proof Complexity

Sarah Sigley
University of Leeds
mmses@leeds.ac.uk

Abstract. In this paper we shall consider the relative proof complexity of three resolution calculi for the modal logic $K_n$, namely $RK_n$ [7] and the calculi of Nalon et al. [14], [15]. The main result of this paper is that all three of these calculi are p-equivalent, despite $RK_n$ being technically rather different to the other calculi. In proving our equivalence result we also introduce a new resolution calculus for $K_n$.

1 Introduction

A multimodal logic over a set of agents $A = \{a_1, \ldots, a_n\}$ is the extension of propositional logic (PL), obtained by adding the set unary operators, $\{\Box a_1, \ldots, \Box a_n\}$. Due to their additional expressiveness such logics are well suited for a variety of applications throughout computer science, as a result there is significant interest in automated theorem proving for the multimodal logic, $K_n$.

Proof complexity measures how efficiently theorems can be proved in a given proof system. Formally, a proof system [5] for some language $L \subseteq \Sigma^*$ is a polynomial time partial function $P : \Sigma^* \rightarrow L$ where $\Sigma^*$ denotes the set of all finite words over $\Sigma$, and a $P$ proof of some $\tau \in L$ is a finite word $\pi \in \Sigma^*$ s.t. $P(\pi) = \tau$. Intuitively, given a proof $\pi$ of a formula $\tau \in L$, an $L$ proof system efficiently verifies that $\pi$ is a correct proof of $\tau$. The size of a proof $\pi$ is the number of symbols it contains, denoted $|\pi|$. Two proof systems are equivalent in terms of their proof complexity if they are $p$-equivalent. That is if given any proof of any formula in the first system we can efficiently construct a proof of the same formula in the second and given any proof of any formula in the second system we can efficiently construct a proof of the same formula in the first.

Determining the complexity of the underlying proof systems of satisfiability algorithms for PL has provided valuable insight into their limitations and guided the development of new algorithms (see [19] for a survey). Recently, work has been carried out concerning the proof complexity of non classical logics e.g. intuitionistic, modal and non-monotonic logics (see [2], [3],[10]). The work that has been carried out on the complexity of $K_n$ proof systems focuses on Frege [8], extended Frege (EF) and substitution Frege (SF) systems [9] for $K_1$. In this paper we will discuss the relative proof complexity of several resolution systems for $K_n$. Propositional resolution is the most well studied proof system for PL, however to our knowledge no previous work has been done concerning the proof complexity of modal resolution systems.
One motivation for studying the proof complexity of modal logics is its connection to the \text{PSPACE} vs. \text{NP} question; it is known that \text{NP} ≠ \text{coNP} iff for every propositional proof system there exists a propositional tautology requiring superpolynomially sized proofs [5]. Analogously, as $K_n$ is \text{PSPACE}-complete [12], \text{PSPACE} ≠ \text{NP} iff for every $K_n$ proof system there exists a $K_n$ tautology requiring superpolynomially sized proofs. Further, certain relationships that hold between proof systems for PL do not hold between the corresponding proof systems for modal logics. For example, the propositional proof systems SF and EF are \text{p-equivalent}, however Jerábek proved in [9] that this is not the case when considering the analogous proof systems $K_1$-SF and $K_1$-EF.

For PL resolution is a very simple calculus consisting of a single inference rule, however constructing an analogous calculus for even the weakest normal modal logic $K_n$ is not straightforward. This is because within a modal formula two complementary literals can occur in different modal contexts and so fail to contradict one another. As a result several different resolution methods have been proposed for $K_n$. In this paper we consider four such clausal resolution systems; three from the literature $RK_n$ [7], $K_n$-Res [14], $K_{ml}$-Res [15] and a new system, $K_{mp}$-Res. The calculus $RK_n$ was one of several direct resolution systems\footnote{A direct resolution system is one that does not involve translation into a more expressive language.} proposed in the late 80's and early 90's (see [13], [6]), since then most proposed resolution methods have required translation into a more expressive language, usually first order logic (see [18] for a survey). The calculi $K_n$-Res and $K_{ml}$-Res are two exceptions to this trend in that they only require translation into a more expressive modal language, not first order logic.

In Sect. 2 we give a brief overview of modal logics (for a full introduction see [4]) followed by a more in depth discussion of resolution calculi, briefly describing propositional resolution and then defining the modal resolution calculi $RK_n$, $K_n$-Res and $K_{ml}$-Res. These modal calculi all operate on different clausal forms, over different modal languages. As a result the proof complexity of the calculus $RK_n$ cannot be directly compared with that of the calculi $K_n$-Res and $K_{ml}$-Res. In Sect. 3 we present our direct $K_n$ resolution calculus, $K_{mp}$-Res. The clausal form that $K_{mp}$-Res operates on is similar to the clausal forms of $K_n$-Res and $K_{ml}$-Res, further, any formula in the clausal form of $K_{mp}$-Res is also in the clausal form of $RK_n$. Consequently the new calculus $K_{mp}$-Res can be directly compared with each of the three other calculi. In Sect. 4 we give our main result; that the resolution calculus $RK_n$, $K_n$-Res, $K_{ml}$-Res and $K_{mp}$-Res are all \text{p-equivalent}, despite the inference rules of the calculus $RK_n$ appearing to be much more complex than the rules of the other calculi on first inspection. Of the calculi we consider only the calculus $K_{ml}$-Res has an associated solver [16], however, choosing to focus practical attention primarily on this simpler system is justified theoretically by our equivalence result. Finally in Sect. 5 we briefly discuss some future work.

Due to space restrictions some proofs are omitted or briefly sketched.
2 Preliminaries

2.1 Modal Logics

A literal is either a propositional variable or its negation, \( \mathcal{L} \) shall denote the set of all literals throughout. We define the modal operator \( \Diamond_a \) s.t. \( \Diamond_a \phi \equiv \neg \square_a \neg \phi \).

A modal literal is either \( \square_a l \) or \( \Diamond_a l \), where \( l \in \mathcal{L} \). In this paper we will take \( \diamond_a \in \{ \square_a, \Diamond_a \} \) and \( \tilde{\mu} = \square_{a_1} \ldots \square_{a_m} \) for \( a = a_1 \ldots a_m \in \mathcal{A}^* \).

The multimodal logic \( \mathbf{K}_n \) is the smallest set containing all propositional tautologies and all formulas \( \mathbf{K}_n : \square_n (\phi \rightarrow \psi) \rightarrow (\square_n \phi \rightarrow \square_n \psi) \), that is closed under the inference rules modus ponens (\( \frac{\phi \phi}{\phi} \)) and \( \alpha \)-necessitation (\( \frac{\phi}{\square \phi} \)).

A Kripke model (henceforth a model) over a set of propositional variables \( P \) and a set of agents \( \mathcal{A} \) is a tuple \( M = (W, R_1, \ldots, R_n, V) \), where \( W \) is a nonempty set of “worlds”, \( R_i \) is a binary relation on \( W \) for each \( a_i \in \mathcal{A} \) and \( V \) is a set of valuation functions \( \{ V(w) : w \in W \} \) s.t. \( V(w) : P \rightarrow \{ \top, \bot \} \).

Let \( \phi, \psi \) be formulas and \( p \in P \). Given a model \( M = (W, R_1, \ldots, R_n, V) \) and a world \( w \in W \) the satisfiability of a formula at \( w \) in \( M \) is defined as follows:

- \((M, w) \models p\) iff \( V(w)(p) = \top \),
- \((M, w) \models \neg \phi\) iff \( (M, w) \not\models \phi \),
- \((M, w) \models \phi \land \psi\) iff \((M, w) \models \phi\) and \((M, w) \models \psi\),
- \((M, w) \models \square_a \phi\) iff \((M, w') \models \phi\) for all \( w' \) s.t. \( (w, w') \in R_a \).

We say a formula \( \phi \) is satisfiable if there exists some world \( w_0 \in W \) in some model \( M \) s.t. \((M, w_0) \models \phi \).

2.2 Modal Resolution Systems

Resolution is a clausal proof system for PL that operates on formulas in conjunctive normal form (CNF). A clause is a disjunction of literals (DoL) and a formula is in CNF if it is a conjunction of clauses. Propositional resolution is a very simple proof system consisting of only a single rule, the resolution rule: \( \frac{C \lor \phi, C' \lor \psi}{C \lor C' \lor \psi \lor \phi} \), where \( C, C' \) are clauses and \( l \in \mathcal{L} \). To prove that a formula is satisfiable using propositional resolution one must convert its negation into CNF and then repeatedly apply the resolution rule until the empty clause is derived.

Clearly any modal formula containing modal operators cannot be translated into CNF. Hence each of the clausal modal resolution systems defined in this section must operate on formulas in some more general modal normal form.

**RK\(_n\)**. In [7] Enjalbert and Fariñas del Cerro proposed a clausal resolution system for \( \mathbf{K}_n \) called \( \mathbf{RK}_n \), which operates on formulas in \( \mathbf{RK}_n \) CNF.

**Definition 1.** A formula \( \phi \) is in \( \mathbf{RK}_n \) conjunctive normal form (\( \mathbf{RK}_n \) CNF) if \( \phi = \bigwedge_{i=1}^N C_i \), where \( C_i \) is an \( \mathbf{RK}_n \) clause. We say \( C \) is an \( \mathbf{RK}_n \) clause if:

\[
C = \ell_1 \lor \cdots \lor \ell_m \lor \bigvee_{x=1}^p (\square_{a_x} D_x) \lor \bigvee_{y=1}^q (\diamond_{a_y} A_y),
\]

where each \( \ell_i \in \mathcal{L} \), each \( D_x \) is an \( \mathbf{RK}_n \) clause and each \( A_y \) is in \( \mathbf{RK}_n \) CNF.
This clausal form is a generalisation of propositional CNF. Note that since an \( \text{RK}_n \) CNF prefixed by \( \Box_a \) can be contained within an \( \text{RK}_n \) clause, a single \( \text{RK}_n \) clause can contain a contradictory pair of literals.

**Definition 2 ([7]).** The rules of the proof system \( \text{RK}_n \) are given in Table 1. The \( \Sigma \) rules are used to compute the resolvent of two \( \text{RK}_n \) clauses whereas the \( \Gamma \) rules are used to compute the resolvent of a single \( \text{RK}_n \) clause.

We say the \( \text{RK}_n \) clause \( C' \) is the unique normal form of \( C \) iff \( C \approx C' \) and no more simplification rules can be applied to \( C' \). Further, \( C \) is a resolvent of the clauses \( C_1, C_2 \), denoted \( \Sigma(C_1, C_2) \Rightarrow C \), iff by repeatedly applying the rules for computing resolvents we can obtain \( \Sigma(C_1, C_2) \Rightarrow C' \) s.t. the normal form of \( C' \) is \( C \). Similarly, we say \( C \) is a resolvent of \( C_1 \) if we can obtain \( \Gamma(C_1) \Rightarrow C \).

<table>
<thead>
<tr>
<th>Rules for Computing Resolvents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A1:</strong> ( \Sigma(p, \neg p) \rightarrow \bot )</td>
</tr>
<tr>
<td><strong>V:</strong> ( \Sigma(A, B) \rightarrow C )</td>
</tr>
<tr>
<td>( \Sigma(A \lor D, B \lor E) \rightarrow C \lor D \lor E )</td>
</tr>
<tr>
<td>( \Sigma(\Box_a A, \Box_a (B \land F)) \rightarrow \Box_a (B \land C \land F) )</td>
</tr>
<tr>
<td><strong>( \Diamond_a: )</strong> ( \Sigma(A, B) \rightarrow C )</td>
</tr>
<tr>
<td>( T(\Box_a (A \land B \land F)) \rightarrow \Box_a (A \land B \land C \land F) )</td>
</tr>
<tr>
<td>( \Gamma(A) \rightarrow B )</td>
</tr>
<tr>
<td>( \Box_a: \Gamma(A) \rightarrow B )</td>
</tr>
<tr>
<td>( \Gamma(\Box_a (A \lor F)) \rightarrow \Box_a (B \lor A \lor F) )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simplification Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S1:</strong> ( \Box_a \bot \approx \bot )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inference rules</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D1:</strong> ( C ) if ( \Gamma(C) \Rightarrow D )</td>
</tr>
</tbody>
</table>

**Example 1.** Let \( \Box_a \mu \) abbreviate the \( \Sigma \) rules \( \Box_a \mu_1, \ldots, \Box_a \mu_n \). We can derive \( \Box_a (\neg l_1 \lor \neg l_2) \) from \( C_1 = \Box_b (\neg l_1 \lor \Box_a l) \) and \( C_2 = \Box_b (\neg l_2 \lor \Box_a l) \) as follows:

| \( \Sigma(l, \neg l) \rightarrow \bot \) | \( \Box_a (\neg l_1 \lor \neg l_2) \rightarrow \bot \) |
| \( \Sigma(\Box_a l, \Box_a \neg l) \rightarrow \bot \lor \Box_a (\bot \land \bot) \) | \( \Sigma(\Box_a l, \Box_a \neg l) \rightarrow \bot \lor \Box_a (\bot \land \bot) \) |
| \( \Sigma(\neg l_1 \lor \Box_a l, \neg l_2 \lor \Box_a \neg l) \rightarrow \neg l_1 \lor \neg l_2 \lor \Box_a (\bot \land \bot) \) | \( \Sigma(\neg l_1 \lor \Box_a l, \neg l_2 \lor \Box_a \neg l) \rightarrow \neg l_1 \lor \neg l_2 \lor \Box_a (\bot \land \bot) \) |

An \( \text{RK}_n \) refutation of the set of \( \text{RK}_n \) clauses \( \{I_1, \ldots, I_n\} \) is a sequence of formulas:
\[ \pi = I_1, \ldots, I_n, R_{(1,1)}, \ldots, R_{(k_1,1)}, C_1, \ldots, C_{m-1}, R_{(1,m)}, \ldots, R_{(k_n,m)}, C_m. \]

Where each \( C_j \) is an RK\(_n\) clause in normal form, \( C_m = \bot \) and for each \( i \in \{2, \ldots, k_j\}, j \in \{1, \ldots, m\} \) either \( R_{(i,j)} \) is an instance of A1 or an instance of A2. Further \( R_{(i,j)} \) is the formula resulting from applying either a rule for computing resolvents or a simplification rule to \( R_{(i-1,j)} \), and \( R_{(k_j,j)} = \Sigma(C, C') \Rightarrow C_j \) or \( F(C) \Rightarrow C_j \), where \( C, C' \in C \cup \{C_1, \ldots, C_{j-1}\} \).

K\(_n\)-Res. More recently, in [14], Nalon and Dixon proposed a clausal resolution system for K\(_n\) which we shall call K\(_n\)-Res. This resolution system determines whether a formula \( \phi \) is satisfiable at some distinguished “start” world, \( s_0 \in W \).

Let \( M = (W, R_1, \ldots, R_n, V) \) be a model and \( w_1, w_2 \in W \). We say \( w_2 \) is reachable from \( w_1 \) if \( (w_1, w_2) \) is in the reflexive transitive closure of \( \bigcup_{i=1}^{n} R_i \).

Note that every world is reachable from itself. We define the master modality, \( \square^* \), s.t. \( (M, w) \models \square^* \phi \iff (M, w') \models \phi \) for all \( w' \) reachable from \( w \).

The clausal form that K\(_n\)-Res operates on is defined below.

**Definition 3 ([14]).** Fix some “start world” \( s_0 \in W \) and let \( S \) be a nullary connective defined s.t. \( (M, w) \models S \) iff \( w = s_0 \). Let \( l, l', l_j \in \mathcal{L} \), a formula \( \phi \) is in separated normal form (SNF) if \( \phi = \bigwedge_{i=1}^{l} C_i \) where each clause \( C_i \) is either:

- Start: \( \square^*(S \rightarrow \bigvee_{j=1}^{l_j} l_j) \),
- Literal: \( \square^*(\bigvee_{j=1}^{l_j} l_j) \),
- Positive modal: \( \square^*(l' \rightarrow \square a l) \),
- Negative modal: \( \square^*(l' \rightarrow \diamond a l) \).

Note that no SNF clause can contain an occurrence of a modal operator with anything more complex than a single literal within its scope. Thus no single clause can be inconsistent, further the number of ways two or more SNF clauses can contradict one another is significantly less than the number of ways two or more RK\(_n\) clauses can do so.

Any formula \( \phi \) in negation normal form\(^2\) (NNF) can be translated into SNF by applying the function \( \tau_0(\phi) = \square^*(S \rightarrow x) \land \tau_1(\square^*(x \rightarrow \phi)) \), where \( x \) is a new propositional variable. Let \( l \in \mathcal{L} \) and \( \phi, \psi \) be formulas then \( \tau_1 \) is defined as follows:

\[
\tau_1(\square^*(x \rightarrow \theta \land \psi)) = \tau_1(\square^*(x \rightarrow \theta) \land \tau_1(\square^*(x \rightarrow \psi)),
\]

\[
\tau_1(\square^*(x \rightarrow o_a \theta)) = \begin{cases} \square^*(x \rightarrow o_a \theta), & \text{if } \theta \in \mathcal{L}, \\ \square^*(x \rightarrow o_a x_1) \land \tau_1(\square^*(x_1 \rightarrow \theta)), & \text{otherwise}. \end{cases}
\]

\[
\tau_1(\square^*(x \rightarrow \theta \lor \psi)) = \begin{cases} \square^*(\neg x \lor \theta \lor \psi), & \text{if } \theta, \psi \text{ DoL's}, \\ \tau_1(\square^*(x \rightarrow \theta \lor x_1)) \land \tau_1(\square^*(x_1 \rightarrow \psi)), & \text{otherwise}. \end{cases}
\]

where \( x_1 \) is a new propositional variable. Intuitively, the function \( \tau_0 \) ties the satisfiability of a formula \( \phi \) to the start world. The function \( \tau_1 \) consists of rewriting

---

\(^2\) A formula over the set of operators \( \{\square, \diamond a, \neg, \land, \lor\} \) where only propositional variables are allowed to be within the scope of a \( \neg \) is in NNF.
and renaming rules; the renaming rules replace subformulas of \( \phi \) with new propositional variables that essentially abbreviate them. A proof that this translation preserves satisfiability is given in [14].

**Definition 4 ([14]).** The inference rules of \( K_n \)-Res are given in Table 2.

\[
\begin{array}{ccc}
\text{IRES1:} & \Box^* (D \lor l) & \text{IRES2:} \Box^* (S \rightarrow D \lor l) & \text{LRES:} \Box^* (D \lor l) \\
\text{MRES:} & \Box^* (t_1 \rightarrow \Box_a l_1) & \Box^* (t_2 \rightarrow \neg \Box_a l_1) \\
\text{GEN2:} & \Box^* (l_1 \rightarrow \Box_a l_1) & \Box^* (l_2 \rightarrow \neg \Box_a l_1) \\
\text{GEN1:} & \Box^* (l_1 \rightarrow \neg \Box_a l_1) & \Box^* (l_2 \rightarrow \Box_a l_1) \\
\text{GEN3:} & \Box^* (l_1 \lor \Box_a l_1) & \Box^* (l_2 \lor \neg \Box_a l_1) \\
\end{array}
\]

Table 2: Rules for \( K_n \)-Res \((l, l', l_j \in L \text{ and } D, D' \text{ are DoL's.})\)

IRES1, IRES2 and LRES correspond to propositional resolution. The rules MRES and GEN2 can be viewed as modal versions of propositional resolution. Note that GEN2 contains \( \Box^* (l_1' \rightarrow \Box_a l_2) \) as \( \Box^* (l_1' \rightarrow \Box_a l_1) \) and \( \Box^* (l_1' \rightarrow \neg \Box_a l_1) \) can both be satisfied by a model \( M \) at a world \( w \in W \) s.t. \((w, w') \notin R_a \) for all \( w' \in W \). The rules GEN1 and GEN3 allow literals to be resolved with modal literals.

**\( K_{ml} \)-Res.** In [15] Nalon, Dixon and Hustadt introduced a layered resolution system for \( K_n \) which we shall call \( K_{ml} \)-Res. This resolution system is similar to \( K_n \)-Res but operates on a labelled clausal form.

**Definition 5 ([15]).** A formula \( \phi \) is in separated normal form with modal levels (SNF\(_{ml}\)) if \( \phi = \bigwedge_{i=1}^{n} C_i \) where each clause \( C_i \) is either:

- **Positive modal:** \( (m : l' \rightarrow \Box_a l_i) \),
- **Literal:** \( (m : \bigvee_{j=1}^{l} l_j) \),
- **Negative modal:** \( (m : l' \rightarrow \neg \Box_a l_i) \),

where \( l, l', l_j \in L \) and \( m \in \mathbb{N} \) represents the modal level\(^3\) of the clause.

A procedure for efficiently translating any NNF formula \( \phi \) into \( \text{SNF}_{ml} \), whilst preserving satisfiability, is given in [15].

**Definition 6 ([15]).** The inference rules of \( K_{ml} \)-Res are given in Table 3.

Labelling clauses by their modal level prevents complementary literals, contained within clauses at different modal levels, from being resolved with one another.

---

\(^3\) A subformula of \( \phi \) has modal level \( m \) iff it is nested within \( m \) modal operators in \( \phi \).
For example, in □ρ form are prefixed by some sequence of modal operators.

However, the clauses obtained by translating some formula φ into □x ∧ aK are prefixed by □ρ.

We shall now present a new resolution system, 3 Resolution with Modal Positions.

3 Resolution with Modal Positions

We shall now present a new resolution system, 3 Resolution with Modal Positions. The primary motivation for defining $K_{mp}$-Res is that it does not require formulas to be translated into some more expressive language and so can be directly compared with other $K_m$ proof systems; as we will see in Sect. 4.

Let φ be a formula. The modal position of a subformula ψ in φ is the finite word $\mu \in A^\ast$ denoting the sequence of modal operators ψ is nested within φ. For example in □q (θl2 ∧ □q l2 ∧ □q (θl3)) the subformulas □q l2 and l3 are at modal positions $a_1 a_2$ and $a_1 a_3 a_1$ respectively.

The normal form that $K_{mp}$-Res operates on is similar to SNF and SNF$_{mt}$. However, the clauses obtained by translating some formula φ into this normal form are prefixed by some sequence of modal operators □μ denoting their modal position within the original formula φ.

Definition 7. Let $l, l', l_j \in L, \mu \in A^\ast$. A formula φ is in separated normal form with modal positions (SNF$_{mp}$) if φ = $\bigwedge_{i=1}^{l} C_i$ where each clause $C_i$ is either:

- Positive modal: $\square_\mu (l' \rightarrow □_\mu l)$
- Negative modal: $\square_\mu (l' \rightarrow \neg □_\mu l)$
- Literal: $\square_\mu (\bigvee_{j=1}^{l} l_j)$
- Negative modal: $\square_\mu (l' \rightarrow \neg □_\mu l)$

To convert an NNF formula φ into SNF$_{mp}$ we apply the translation $T(\phi) = x \land \rho(\square_\varepsilon (x \rightarrow \phi))$, where x is a new propositional variable, ε denotes the empty word and ρ is defined as follows. Let $x_1$ be a new propositional variable, $\mu \in A^\ast$:

- $\rho(\square_\mu (x \rightarrow \theta \land \psi)) = \rho(\square_\mu (x \rightarrow \theta)) \land \rho(\square_\mu (x \rightarrow \psi))$
- $\rho(\square_\mu (x \rightarrow \varnothing_\theta)) = \left\{ \begin{array}{ll} \square_\mu (x \rightarrow \varnothing_\theta) & \text{if } \theta \in L, \\ \square_\mu (x \rightarrow \varnothing_\theta \land \rho(\square_\mu (x \rightarrow \psi)) & \text{if } \theta, \psi \text{ DoL's,} \end{array} \right.$

Table 3. Rules for $K_{mp}$-Res ($l, l', l_j \in L, m \in N, D, D' \text{ are DoL's}$.)

3 Resolution with Modal Positions

We shall now present a new resolution system, 3 Resolution with Modal Positions. The primary motivation for defining $K_{mp}$-Res is that it does not require formulas to be translated into some more expressive language and so can be directly compared with other $K_m$ proof systems; as we will see in Sect. 4.

Let φ be a formula. The modal position of a subformula ψ in φ is the finite word $\mu \in A^\ast$ denoting the sequence of modal operators ψ is nested within φ. For example in □q (θl2 ∧ □q l2 ∧ □q (θl3)) the subformulas □q l2 and l3 are at modal positions $a_1 a_2$ and $a_1 a_3 a_1$ respectively.

The normal form that $K_{mp}$-Res operates on is similar to SNF and SNF$_{mt}$. However, the clauses obtained by translating some formula φ into this normal form are prefixed by some sequence of modal operators □μ denoting their modal position within the original formula φ.

Definition 7. Let $l, l', l_j \in L, \mu \in A^\ast$. A formula φ is in separated normal form with modal positions (SNF$_{mp}$) if φ = $\bigwedge_{i=1}^{l} C_i$ where each clause $C_i$ is either:

- Positive modal: $\square_\mu (l' \rightarrow □_\mu l)$
- Negative modal: $\square_\mu (l' \rightarrow □_\mu l)$
- Literal: $\square_\mu (\bigvee_{j=1}^{l} l_j)$
- Negative modal: $\square_\mu (l' \rightarrow □_\mu l)$

To convert an NNF formula φ into SNF$_{mp}$ we apply the translation $T(\phi) = x \land \rho(\square_\varepsilon (x \rightarrow \phi))$, where x is a new propositional variable, ε denotes the empty word and ρ is defined as follows. Let $x_1$ be a new propositional variable, $\mu \in A^\ast$:

- $\rho(\square_\mu (x \rightarrow \theta \land \psi)) = \rho(\square_\mu (x \rightarrow \theta)) \land \rho(\square_\mu (x \rightarrow \psi))$
- $\rho(\square_\mu (x \rightarrow \varnothing_\theta)) = \left\{ \begin{array}{ll} \square_\mu (x \rightarrow \varnothing_\theta) & \text{if } \theta \in L, \\ \square_\mu (x \rightarrow \varnothing_\theta \land \rho(\square_\mu (x \rightarrow \psi)) & \text{if } \theta, \psi \text{ DoL's,} \end{array} \right.$
Proposition 1. An NNF formula $\phi$ is satisfiable iff $T(\phi) = x \land \rho(\square_c(x \to \phi))$ is satisfiable.

Proof. ($\Rightarrow$): Let $M = (W, R_1, \ldots, R_n, V)$ be a model s.t. $(M, w_0) \models \phi$ for some $w_0 \in W$. We let $M_1 = (W, R_1, \ldots, R_n, V_1)$ where $V_1(w_0)(x) = \perp$, $V_1(w_0)(x) = \top$, and for all $w \in W$ and all variables $p$ in the domain of $V(w)$ we have $V_1(w)(p) = V(w)(p)$. Then $(M_1, w_0) \models \phi$ and so $(M_1, w_0) \models x \to \phi$. We will prove by induction on the structure of $\phi$ that there exists a model $M_2 = (W, R_1, \ldots, R_n, V_2)$ s.t. $(M_2, w_0) \models \rho(\square_c(x \to \phi))$ and for all $w \in W$ and all propositional variables $p$ in the domain of $V_1(w)$ we have $V_2(w)(p) = V_1(w)(p)$. From which it follows immediately that $(M_2, w_0) \models x$ and so $(M_2, w_0) \models x \land \rho(\square_c(x \to \phi))$.

There are two base cases for the induction; the case where $\phi$ is a modal literal and the case where $\phi$ is a DoL. In both cases it follows immediately from the definition of $\rho$ that $(M_1, w_0) \models \rho(\square_c(x \to \phi))^4$. Hence we take $M_2 = M_1$.

Suppose $\phi = \theta \lor \psi$ and at least one of $\theta$ and $\psi$ is not a DoL. Then $\rho(\square_c(x \to \phi)) = \rho(\square_c(x \to \phi)) \land \rho(\square_c(x \to \phi))$. Let $M_2 = (W, R_1, \ldots, R_n, V_2)$ where $V_2(w)(p) = V_1(w)(p)$ for all variables $p$ in the domain of $V_2$, $V_2(w_m)(x_1) = \top$ for all $w_m \in W$ s.t. $(M_1, w_m) \models \psi$ and $w_m$ is $\mu$-reachable from $w_0$, and $V_2(w')(x_1) = \perp$ for all other $w' \in W$. Then $(M_2, w_0) \models \square_c(x_1 \to \psi)$ and $(M_2, w_0) \models \square_c(x \to \theta \lor x_1)$. By the inductive hypothesis there exists models $M'_1 = (W, R_1, \ldots, R_n, V'_1)$ s.t. $(M'_1, w_0) \models \rho(\square_c(x \to \theta \lor x_1))^6$ and $M'_2 = (W, R_1, \ldots, R_n, V'_2)$ s.t. $(M'_2, w_0) \models \rho(\square_c(x_1 \to \psi))$. Further $M'_1(w)(p) = M'_2(w)(p)$ for all $w \in W$ and all $p$ in the domain of $V_1$. For $i \in \{1, 2\}$ let $X_i$ be the domain of $V'_i(w_0)$, then for all $w \in W$ let $V_3(w)(p) = V'_i(w)(p)$ if $p \in X_i$ and $V_3(m)(p) = V'_2(w)(p)$ if $p \in X_2 \setminus X_1$. Then $(M_3, w_0) \models \rho(\square_c(x \to \theta \lor x_1)) \land \rho(\square_c(x_1 \to \psi))$.

The cases for $\phi = \diamond_c \theta$ and $\phi = \theta \land \psi$ are similar.

($\Leftarrow$): If $(M, w_0) \models x \land \rho(\square_c(x \to \phi))$ then $(M, w_0) \models \rho(\square_c(x \to \phi))$ and $(M, w_0) \models x$. We will prove by induction on the structure of $\phi$ that $(M, w_0) \models (x \to \phi)$ from which it follows that $(M, w_0) \models \phi$.

For the base cases we suppose $\phi$ is either a modal literal or a DoL. In both cases $(M, w_0) \models \square_c(x_m \to \phi)$ by the definition of $\rho$. Here $x_m$ is a variable that was added by the translation $\rho$ and is satisfied at every $w_m \in W$ s.t. $w_m$ is $\mu$-reachable from $w_0$, note that if $\mu = \varepsilon$ then $x_m = x$. If $\phi = \square \theta \land \psi$, then $\rho(\square_c(x_m \to \phi)) = \rho(\square_c(x_m \to \theta \land \psi))$ and so by assumption $(M, w_0) \models \rho(\square_c(x_m \to \theta) \land \rho(\square_c(x_m \to \psi)))$. Hence the inductive hypothesis $(M, w_0) \models \square_c(x_m \to \theta)$ and $(M, w_0) \models \square_c(x_m \to \psi)$ and so $(M, w_0) \models \square_c(x_m \to \theta \land \psi)$. The cases for $\phi = \theta \land \psi$ and $\phi = \diamond_c \theta$ are similar.

\[ \square \text{ Note that here we have replaced } \square_c \text{ with } \square_m \text{ where } \mu \text{ is some arbitrary finite word in } \mathcal{A}^* \text{ (possibly } \varepsilon). \text{ This is necessary to apply the inductive hypothesis as for instance, if } \phi = \square_a D \text{ where } D \text{ is a DoL then } \rho(\square_c(x \to \phi)) = \square_c(x \to \square_a x_1) \land \rho(\square_c(x_1 \to D)). \]

\[ \text{A world } w_m \text{ is } \mu = a_1, a_2, \ldots, a_m \text{-reachable from } w_0 \text{ if there exists a path } (w_0, w_1, \ldots, (w_{m-1}, w_m) \text{ from } w_0 \text{ to } w_m \text{ where } (w_{i-1}, w_i) \in R_i. \text{ Further } w_m \text{ is } \varepsilon \text{-reachable from } w_0 \text{ iff } w_m = w_0. \]

\[ \begin{multline*} \text{Since } \rho(\square_c(x \to \theta \lor x_1)) = \rho(\square_c(x \to (x_1 \lor x_2))) \land \rho(\square_c(x_2 \to \theta)). \end{multline*} \]
Definition 8. The inference rules of $K_{mp}$-Res are given in Table 4.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LRES: $\Box_\mu(D \lor l)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plateau $\Box_\mu(D' \lor \lnot l)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRES: $\Box_\mu(l_1 \rightarrow \Box_\mu l_1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(l_2 \rightarrow \lnot_\mu l_1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(\lnot l_1 \lor \lnot l_2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEN2: $\Box_\mu(l_1' \rightarrow \Box_\mu l_1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(l_2' \rightarrow \Box_\mu l_2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(\lnot l_1' \lor \lnot l_2' \lor \lnot l_3')$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEN1: $\Box_\mu(l_1' \rightarrow \Box_\mu l_1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(l_2' \rightarrow \Box_\mu l_2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(l_3' \rightarrow \Box_\mu l_3)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(l_1' \lor \cdots \lor l_2' \lor l_3')$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEN3: $\Box_\mu(l_1' \rightarrow \Box_\mu l_1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(l_2' \rightarrow \Box_\mu l_2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(l_3' \rightarrow \Box_\mu l_3)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Box_\mu(l_1' \lor \cdots \lor l_2' \lor l_3')$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Rules for $K_{mp}$-Res ($l, l', l_j \in L$, $\mu \in A^*$, $D, D'$ are DoL's.)

Theorem 1. $K_{mp}$-Res terminates and is sound and complete.

Proofs of soundness and termination for $K_{mp}$-Res follow directly from the respective proofs for $K_{ml}$-Res given in [15]. A proof of completeness can also be found by adapting the analogous proof in [15].

4 Simulations

We can compare the strength of two $K_n$ proof systems via p-simulations.

Definition 9. Let $Q_1, Q_2$ be $K_n$ proof systems. We say $Q_1$ p-simulates $Q_2$, denoted $Q_1 \geq_p Q_2$, iff there exists some polynomial time function, $f$, s.t. for all $\tau \in K_n$ and all $Q_2$ proofs, $\pi$, of $\tau$ the word $f(\pi)$ is a $Q_1$ proof of $\tau$. If we also have $Q_1 \leq_p Q_2$ then $Q_1$ and $Q_2$ are p-equivalent, denoted $Q_1 \equiv_p Q_2$.

We will show that the systems $K_n$-Res, $K_{ml}$-Res and $K_{mp}$-Res are p-equivalent. Let $\phi$ be a formula in NNF and $C$ be the corresponding set of SNF clauses. We can keep track of the modal position of each clause in $C$ by labelling the extension variables$^7$ of $C$ by two finite words, $\mu \in A^*$ and $\lambda \in N^*$.$^8$

Let $l \in L$ and let $\phi, \theta$ and $\psi$ be formulas. We describe this labelling process by redefining the translation functions $\tau_0$ and $\tau_1$ so that they contain labelled extension variables as follows:

$^7$ The variables introduced in the translation $\tau_0(\phi)$.

$^8$ Denoting the order the extension variables at modal position $\mu$ were introduced.
\[\tau_0^\bullet(\phi) = \Box^*(S \rightarrow x_\xi^\epsilon) \land \tau_1^\bullet(\Box^*( x_\xi^\epsilon \rightarrow \phi)),\]
\[\tau_1^\bullet(\Box^*(x_\mu^\lambda \rightarrow \theta \land \psi)) = \tau_1^\bullet(\Box^*( x_\mu^\lambda \rightarrow \theta)) \land \tau_1^\bullet(\Box^*( x_\mu^\lambda \rightarrow \psi)),\]
\[\tau_1^\bullet(\Box^*(x_\mu^\lambda \rightarrow o_\alpha\theta)) = \begin{cases} 
\Box^*( x_\mu^\lambda \rightarrow o_\alpha\theta), & \text{if } \theta \in \mathcal{L}, \\
\Box^*( x_\mu^\lambda \rightarrow o_\alpha x_1) \land \tau_1^\bullet(\Box^*( x_\mu^\lambda \rightarrow \theta)), & \text{otherwise}. 
\end{cases}\]
\[\tau_1^\bullet(\Box^*(x_\mu^\lambda \rightarrow \theta \lor \psi)) = \begin{cases} 
\Box^*( x_\mu^\lambda \lor \theta \lor \psi), & \text{if } \theta, \psi \text{ DoL's,} \\
\tau_1^\bullet(\Box^*( x_\mu^\lambda \rightarrow \theta \lor x_1)) \land \tau_1^\bullet(\Box^*( x_\mu^\lambda \rightarrow \psi)), & \text{otherwise}. 
\end{cases}\]

where \( N = 0 \) if there is no \( M \in \mathbb{N} \) s.t. \( x_\mu^{\lambda M} \in \mathcal{X} \), otherwise \( N = \max\{\mathcal{M}_1 \in \mathbb{N} : x_\mu^{\lambda M_1} \in \mathcal{X}\} + 1 \). Here \( \mathcal{X} \) is the set of previously introduced extension variables.

Let \( \pi \) be a refutation of some set of SNF clauses \( \mathcal{C} \) and let \( C, C' \) be clauses in \( \pi \). We say \( C' \) is a descendant of \( C \) if it was derived by applying an inference rule to a set of clauses containing either \( C \) or one of its descendants.

**Lemma 1.** Let \( \phi \) be a formula, let \( \mathcal{C} = \tau_0^\bullet(\phi) \) and let \( \pi \) be a \( K_n \)-Res refutation of \( \mathcal{C} \). Let \( C \) be a clause in \( \pi \) derived either by applying IRES1 or IRES2 to a pair of clauses at different modal positions or by applying some other inference rule of \( K_n \)-Res to some set of clauses whose modal positions do not agree with those of Definition 8. The sequence of clauses obtained by deleting every such \( C \) and all of its descendants from \( \pi \) is also a \( K_n \)-Res refutation of \( \mathcal{C} \).

**Proof.** For each inference rule of \( K_n \)-Res we can show that if any such clause \( C \) was derived using said rule then \( \bot \) is not a descendant of \( C \). It follows that every such \( C \) and its descendants can be deleted from \( \pi \) to obtain new refutation.

(LRES): By the definition of \( \tau_0^\bullet \) the two clauses \( C_1 \) and \( C_2 \) from which \( C \) was derived must each contain a negative instance of some extension variable, say \( \neg x_\mu^{\lambda_1} \) and \( \neg x_\mu^{\lambda_2} \) respectively. Hence \( C \) must also contain these literals. Suppose \( \bot \) is a descendant of \( C \), then by induction on the lengths of \( \lambda_1 \) and \( \lambda_2 \), and the definition of \( \tau_0^\bullet \) we can show that there exists some \( C' \) which is a descendant of \( C \) s.t. \( \bot \) is a descendant of \( C' \) and \( C' \) contains the literals \( \neg x_\mu^{\lambda_1} \) and \( \neg x_\mu^{\lambda_2} \). However \( x_\mu^{\lambda_1} \) and \( x_\mu^{\lambda_2} \) only occur positively in \( C \), and so \( \pi \), (by Definition 4) as modal literals. To derive \( \bot \) from \( C' \) either GEN1 or GEN3 must be applied to some set of clauses containing either \( C' \) or one of its descendants in which \( \neg x_\mu^{\lambda_1} \) and \( \neg x_\mu^{\lambda_2} \) appear. It can be shown that no such set exists. The cases for the other rules are similar. \( \square \)

**Theorem 2.** \( K_n \)-Res \( \equiv_p \) \( K_{ml} \)-Res \( \equiv_p \) \( K_{mp} \)-Res.

**Proof.** Let \( \phi \) be an NNF formula. Translating \( \phi \) into SNF, SNF_{ml} and SNF_{mp} we obtain three sets of clauses, \( \mathcal{C}, \mathcal{C}_{ml} \) and \( \mathcal{C}_{mp} \). There is a one to one correspondence between the clauses in each set, where for any \( \mu \in \mathcal{A}^* \) s.t. \( |\mu| = m \):
\[\tilde{\mu}_0(D \lor \neg x_\mu^{\lambda_1}) \in \mathcal{C}_{mp} \iff (m : D \lor \neg x_\mu^{\lambda_1}) \in \mathcal{C}_{ml} \iff \Box^*(\neg x_\mu^{\lambda_1} \lor D) \in \mathcal{C},\]
\[\tilde{\mu}_0(x_\mu^\lambda \rightarrow \Box_a l) \in \mathcal{C}_{mp} \iff (m : x_\mu^\lambda \rightarrow \Box_a l) \in \mathcal{C}_{ml} \iff \Box^*(x_\mu^\lambda \rightarrow \Box_a l) \in \mathcal{C},\]
\[\tilde{\mu}_0(x_\mu^\lambda \rightarrow \Box_a l) \in \mathcal{C}_{mp} \iff (m : x_\mu^\lambda \rightarrow \Box_a l) \in \mathcal{C}_{ml} \iff \Box^*(x_\mu^\lambda \rightarrow \Box_a l) \in \mathcal{C},\]
\[x_\xi^\epsilon \in \mathcal{C}_{mp} \iff (0 : x_\xi^\epsilon) \in \mathcal{C}_{ml} \iff \Box^*(S \rightarrow x_\xi^\epsilon) \in \mathcal{C}.\]
Corollary 1. Let $\pi_{mp}$ be a $K_{mp}$-Res refutation of $C_{mp}$. If we take $\pi_{ml}$ and $\pi$ to be the corresponding sequences of SNF mp and SNF clauses then we obtain a $K_{ml}$-Res refutation of $C_{ml}$ and a $K_n$-Res refutation of $C$.

($\geq p$): Let $\pi_{mp}$ be a $K_{mp}$-Res refutation of $C_{mp}$. Suppose that $\pi_{ml}$ and $\pi$ are the corresponding sequences of SNF mp and SNF clauses then we obtain a $K_{ml}$-Res refutation of $C_{ml}$ and a $K_n$-Res refutation of $C$.

($\leq p$): Now suppose $\pi$ is a $K_n$-Res refutation of $C$. Let $\pi'$ be the refutation obtained by applying Lemma 1 to $\pi$ and let $\pi_{ml}'$ and $\pi_{mp}'$ be the corresponding sequences of SNF mp and SNF clauses. Then $\pi_{ml}'$ and $\pi_{mp}'$ are refutations of $C_{ml}$ and $C_{mp}$ respectively.

Finally we compare the proof systems $K_{mp}$-Res and $RK_n$. We only need to consider refutations of sets of unsatisfiable SNF mp clauses as every such set is also an unsatisfiable formula in $RK_n$ CNF and every $RK_n$ CNF can be translated into SNF mp. However note that an $RK_n$ refutation of a set of SNF mp clauses can contain $RK_n$ clauses that are not also SNF mp clauses.

Theorem 3. $K_{mp}$-Res $\equiv_p RK_n$.

Proof. ($\leq p$): We can prove that the proof system $K_{mp}$-Res is $p$-simulated by the proof system $RK_n$ by showing that each of $K_{mp}$-Res's inference rules can be $p$-simulated using the rules of $RK_n$. In Example 1 we gave a simulation for MRES. The simulations for the other rules are similar.

($\geq p$): Let $\pi$ be an $RK_n$ refutation of some set of SNF mp clauses $C$ and let $\mathcal{MP}$ be the set of modal positions of the clauses in $C$. We can translate $\pi$ into a layered normal form. This normal form consists of $|\mathcal{MP}|$ blocks of $RK_n$ clauses and their associated formulas for computing resolvents. The first block contains only $RK_n$ clauses at some $\mu_1 \in \mathcal{MP}$ of maximal length, the second block contains $RK_n$ clauses at $\mu_2 \neq \mu_1 \in \mathcal{MP}$ s.t. $|\mu_2| \leq |\mu_1|$ and so on.

Further within each block there are five subblocks corresponding to the rules GEN1, GEN3, MRES, GEN2 and LRES respectively. Consider the block of $RK_n$ clauses at position $\mu$. The first subblock contains $RK_n$ clauses in groups of $z+1$, where the first of these $RK_n$ clauses was derived by resolving some $RK_n$ clause $\Box_\mu(l_1' \rightarrow \Box_a \neg l_1)$ with some $RK_n$ clause $\Box_\mu(l_1 \lor \cdots \lor l_{z+1})$. For each $j \in \{2, \ldots, z\}$ the $j$th $RK_n$ clause was derived by resolving the $j$-th $RK_n$ clause with $\Box_\mu(l_j' \rightarrow \Box_a \neg l_j)$. The $z+1$th $RK_n$ clause is $\Box_\mu(\neg l_1' \lor \cdots \lor \neg l_{z+1})$ obtained by resolving the $z$th $RK_n$ clause with some negative modal SNF mp clause $\Box_\mu(l_{z+1} \rightarrow \Diamond_a \neg l_{z+1})$. The $RK_n$ clauses in the second block are similarly in groups of $z+1$ where the first $RK_n$ clause was derived by resolving $\Box_\mu(l_1 \lor \cdots \lor l_z)$ with $\Box_\mu(l_1' \rightarrow \Box_a \neg l_1)$ and every other $RK_n$ clause was derived as in the first block. The third subblock contains only $RK_n$ clauses derived by resolving a modal literal with its negation. The $RK_n$ clauses within the fourth subblock are in pairs s.t. the first $RK_n$ clause is of the form $\Box_\mu(\neg l_1 \lor \neg l_2 \lor l_a \bot)$ and was derived by resolving some $RK_n$ clause $\Box_\mu(l_1 \rightarrow \Box_a \bot)$ with $\Box_\mu(l_2 \rightarrow \Box_a \neg l)$ and the second $RK_n$ clause was obtained by resolving the first with some $\Box_\mu(l_3 \rightarrow \Diamond_a l)$. The $RK_n$ clauses in the fifth subblock were obtained by resolving a literal and its negation. The size of this layered refutation is polynomial in $|\pi|$ and from it we can obtain a $K_{mp}$-Res refutation of $C$. 

Corollary 1. $K_n$-Res $\equiv_p K_{ml}$-Res $\equiv_p K_{mp}$-Res $\equiv_p RK_n$. 

70
5 Future Work

We have shown that several resolution systems for the modal logic $K_n$ are equivalent in terms of their proof complexity. A natural extension of this work would be to determine whether any of the standard techniques for obtaining lower bounds in proof complexity can be applied to the modal resolution systems in this paper. In particular feasible interpolation [11], [17] and size-width [1] are two such techniques that have yielded exponential lower bounds for propositional resolution and so would be obvious candidates.

References

Might Counterfactual Donkey Sentences

Sam Carter and Simon Goldstein
Rutgers University, New Brunswick NJ 08904, USA

1 Introduction

This paper explores donkey sentences that are *might* counterfactuals, like (1):

(1) If an animal had escaped from the zoo, it might have attacked Alex.

Donkey *might* counterfactuals have a layer of ambiguity not found in ordinary donkey counterfactuals, like (2):

(2) If an animal had escaped from the zoo, it would have attacked Alex.

It is well known that donkey counterfactuals exhibit an ambiguity between ‘selective’ and ‘unselective’ readings (van Rooij 2006; Wang 2009; Walker and Romero 2015). Since we are interested in donkey anaphora, set aside the reading on which an *animal* takes wide scope over the conditional. Two readings of (2) are then available. The unselective reading says that for every animal, *d*, if *d* had escaped from the zoo, it would have attacked Alex. The selective reading quantifies over a smaller set of animals. It concerns the set of animals which escape from the zoo in one of the nearest worlds where some animal escapes. Or, in other words, the set of animals which might have escaped had any animal escaped. The selective reading says that for every one of those animals, *d*, if *d* had escaped, it would have attacked Alex. So let *D* be the set of all animals in the zoo, and *D* near be the subset of those animals which escape in one of the nearest ‘escaping animal’-worlds. Then we have the following two readings:

(2a) For every *d* ∈ *D*: if *d* had escaped, *d* would have attacked Alex. \textit{unselective}

(2b) For every *d* ∈ *D* near: if *d* had escaped, *d* would have attacked Alex. \textit{selective}

Our thesis is that, in addition to the selective / unselective ambiguity, *might*-donkey counterfactuals exhibit a further, orthogonal ambiguity between what we will call ‘universal’ and ‘existential’ readings. This extra ambiguity is not present in ordinary donkey counterfactuals. Selective and unselective readings differ in the domain of quantification. By contrast, existential and universal readings differ in quantificational force. The universal reading of (1) says that for every *d* in the relevant domain, if *d* had escaped, it might have attacked Alex. By contrast, the existential reading says that there is some *d* in the relevant domain where if *d* had escaped, it might have attacked Alex.

This generates four readings of (1), ordered by decreasing strength:
(1) a. For every $d \in D$: if $d$ had escaped, $d$ might’ve attacked Alex. 
UNSELECTIVE UNIVERSAL

b. For every $d \in D_{\text{near}}$: if $d$ had escaped, $d$ might’ve attacked Alex. 
SELECTIVE UNIVERSAL

c. For some $d \in D_{\text{near}}$: if $d$ had escaped, $d$ might’ve attacked Alex. 
SELECTIVE EXISTENTIAL

d. For some $d \in D$: if $d$ had escaped, $d$ might’ve attacked Alex. 
UNSELECTIVE EXISTENTIAL

To individuate these readings, let’s consider the range of cases in which each is true. So suppose there are two animals in the zoo: $a$ (ndy the Alligator) and $b$ (esty the Bunny). $b$ wouldn’t attack Alex, but $a$ might. If an animal’s gate is open, the possibility of that animal escaping is closer than the possibility of any animal escaping whose gate is closed. We can consider three scenarios which validate successively stronger readings of (1):

(3) i. The zookeeper leaves only the gate to $b$’s cage open.
   ii. The zookeeper leaves the gate to $a$ and $b$’s cages open.
   iii. The zookeeper leaves only the gate to $a$’s cage open.

Only the Unselective Existential reading of (1) is true in (3i), since in all the nearest worlds in which an animal escapes, it is $b$. The Selective and Unselective Existential readings are true (3ii), since it contains a nearest animal escaping world where $a$ escapes. Only the Unselective Universal reading is false in (3iii), since in all the nearest worlds in which an animal escapes, it is $a$. Finally, every reading is false if $b$ is the only animal in the zoo, and every reading is true if $a$ is the only animal in the zoo.

In this paper, we explain the pattern of readings above by integrating the theory of anaphora and possibility modals from Groenendijk et al. 1996 with Starr 2014’s account of counterfactuals. The resulting theory has several upshots, both empirical and theoretical. First, we provide a new analysis of the selective / unselective ambiguity. This is derived as a structural ambiguity between the relative scope of the indefinite and subjunctive mood. Second, we provide an analysis of the existential / universal ambiguity. This ambiguity arises through the ability of possibility modals to operate in two different ways. The possibility modal can either be a global test on a context, requiring that the context can be consistently updated with its prejacent. Or, alternatively, the possibility modal can update the context, filtering out any potential values of variables that cannot survive update with the prejacent.

Ultimately, our explanation of both ambiguities relies on uncovering a previously unnoticed structural property of dynamic semantics. We show that the semantics for indefinites originally introduced in Groenendijk et al. 1996 can be simplified and restated through the introduction of a special quasi-distributive operator. The key idea is that when contexts contain information about both assignments and worlds, it is important that this information be packaged in the right way. In particular, we introduce an operation that partitions contexts (sets
of assignment-world pairs) into cells that agree on the assignment. Each cell is then updated, and the results unioned together. We state the meaning of both indefinites and possibility modals as the composition of this quasi-distributive operator with several other operators. Both ambiguities that we address are explained through this basic idea. Ultimately, each of these readings correspond to a different choice in whether to test a context one cell at a time, or globally.

2 Indicative & Subjunctive Conditionals

The conditionals in (4) and (5) differ in meaning.

(4) If the alligator \( x \) escaped from the zoo, it \( x \) has returned.

(5) If the alligator \( x \) had escaped from the zoo, it \( x \) would have returned.

Suppose that the speaker is looking at the alligator in its cage. Under these conditions, an utterance of (4) is true, but (5) might be false. In addition to differing in their truth conditions, (4) and (5) differ regarding the contexts in which they can be appropriately asserted. If it is common knowledge that the alligator did not escape from the zoo, then an utterance of (4) is marked. By contrast, an utterance of (5) remains felicitous (even if false). (5), but not (4), can be used counterfactually.

Call conditionals of the former kind 'indicative' conditionals, and conditionals of the latter kind 'subjunctive'. This distinction is ubiquitous (for example: Stalnaker 1968; Adams 1970; Lewis 1973; and Gibbard 1981). Yet there is no consensus on whether it should be explained in syntactic, semantic or pragmatic terms. Starr 2014, following Stalnaker 1975, defends a uniform analysis of indicative and subjunctive conditionals, which assigns the same meaning to the conditional operator (\( \rightarrow \)) in both cases.

On this analysis, the contrast between (4) and (5) is explained by grammatical mood. Instead of expressing tense, Starr (simplifying Schulz 2007) suggests that the morphology in the antecedent of (5) functions to encode subjunctive mood (\( /lhd \)), which scopes below the antecedent of the conditional, modeled with \( \rightarrow \). Indicative and subjunctive conditionals like (4) and (5) are then ascribed the forms \( \phi \rightarrow \psi \) and \( /lhd \phi \rightarrow \psi \).

3 Update Semantics & Subjunctive Mood

In update semantics, the meaning assigned to a sentence reflects not just its truth conditions but also its effect on a context.

Definition 1. \( \mathcal{L} \) is a set of atomic sentences \( \mathcal{L}_a \), closed under \( \neg, \Diamond, \Box, \rightarrow, \) and \( /lhd \). A possible world \( w \) is a function from atomic sentences to truth values. \( W \) is the set of such \( w \)’s. A context \( c \) is a set of worlds. An interpretation function \([\cdot]\) maps sentences to context change potentials, functions from contexts to contexts. A selection function \( f \) maps a world \( w \) and a set of worlds \( A \) to the set of closest worlds to \( w \) where \( A \) is true. A model \( M = (W, [\cdot], f) \), for some \( W, [\cdot], \) and \( f \).
We define entailment in terms of context change. Say that a context $c$ supports $\phi$ ($c \models \phi$) iff $\phi$ has no effect on $c$. Then some premises entail a conclusion iff any context supports the conclusion after having been updated with the premises.

**Definition 2.**

1. $c$ supports $\phi$ ($c \models \phi$) iff $c[\phi] = c$.
2. $\phi_1; \ldots; \phi_n \models \psi$ iff for any context $c$, $c[\phi_1] \ldots [\phi_n] \models \psi$.

We follow Veltman 1996 and Gillies 2004 in interpreting the fragment of $\mathcal{L}$ without $\triangleleft$.

**Definition 3.**

1. $c[\alpha] = \{ w \in c \mid w(\alpha) = 1 \}$
2. $c[\neg \phi] = \{ w \in c \mid w \not\models \phi \}$
3. $c[\Diamond \phi] = \{ w \in c \mid c[\phi] \neq \emptyset \}$
4. $c[\Box \phi] = \{ w \in c \mid c = \emptyset \}$
5. $c[\phi \rightarrow \psi] = \{ w \in c \mid c[\phi] \models \psi \}$

Atoms narrow down a context to the worlds where they are true. $\neg \phi$ collects every world from $c$ that would not survive updating with $\phi$. $\Diamond \phi$, $\Box \phi$, and $\phi \rightarrow \psi$ are tests. They either leave the context the same, or produce the absurd state $\emptyset$.

$\Diamond \phi$ tests whether updating with $\phi$ is consistent. $\Box \phi$ tests whether $\phi$ is supported. $\phi \rightarrow \psi$ tests whether updating with $\phi$ creates a state in which $\psi$ is supported (Gillies 2004).

Following Starr 2014, we model subjunctive conditionals with a mood operator ($\triangleleft$). $\triangleleft \phi$ expands a context by finding the closest worlds where the proposition associated with $\phi$ is true. So our models contain a selection function $f$.

We assume that the selected worlds always make the inputted proposition true, and that whenever the inputted proposition is true at the inputted world, the inputted world is the unique output.

**Definition 4.**

1. $f(w, [\phi]) \subseteq [\phi]$.
2. If $w \in [\phi]$, then $f(w, [\phi]) = \{ w \}$.

So let $[\phi]$ denote the set of worlds $w$ which themselves support $\phi$. Then $\triangleleft \phi$ is a pointwise update that finds the closest worlds to each $w \in c$ where $[\phi]$ is true, collecting the results.

**Definition 5.**

1. $[\phi] = \{ w \mid w \models \phi \}$
2. $c[\triangleleft \phi] = \bigcup \{ f(w, [\phi]) \mid w \in c \}$

$\triangleleft \phi$ is a non-eliminative update; where $c \nsubseteq [\phi]$, it is not guaranteed that $c[\triangleleft \phi] \subseteq c$. The subjunctive conditional is analyzed as $\triangleleft \phi \rightarrow \psi$.

This semantics explains the difference between indicative and subjunctive conditionals. Where $\phi$ does not contain $\triangleleft$, $\phi \rightarrow \psi$ takes the subset of $c$ which
survives update with $[\phi]$ and checks that it supports $\psi$. By contrast, $\langle \phi \to \psi \rangle$ takes that subset, plus the set of nearest $[\phi]$-worlds to the context, and checks that this set supports $\psi$.

To see how indicatives and subjunctives function differently, return to (4). The context described in §1 supports (4). That context supports the claim that the alligator is in the cage. By contrast, there is no guarantee that in some contextually live possibility in which the alligator did not escape, the nearest possibilities in which it escaped are possibilities in which it returned. So, the context need not support (5).

This analysis also explains why indicatives like (4) can’t be used counterfactually. Where $c[\phi] = \emptyset$, $c[\phi]$ trivially supports $\psi$. A prohibition on trivial tests then predicts the infelicity of uttering (4) in contexts in which the alligator is known not to have escaped. By contrast, where $c[\phi] = \emptyset$, $\langle \phi \to \psi \rangle$ is a non-trivial test on $c$.

4 Donkey Conditionals

Donkey conditionals such as (6) feature an indefinite scoping inside the antecedent, co-indexed with a pronoun in the consequent.

(6) If an animal $x$ escaped from the zoo, the zoo-keepers found it $x$.

(6) exhibits two noteworthy features common to donkey conditionals (Geach 1962). First, the denotation of the pronoun in the consequent co-varies with the indefinite in the antecedent, despite occurring outside of its scope. Second, the conditional has a universal interpretation: it entails that, for all animals $x$, if $x$ escaped from the zoo, the zoo-keepers found $x$.

Donkey conditionals can also occur with subjunctive mood:

(7) If an animal $x$ had escaped from the zoo, the zoo-keepers would have found it $x$.

(8) If an animal $x$ had escaped from the zoo, the zoo-keepers might have found it $x$.

In order to account for the two features of donkey conditionals described above, we combine the treatment of subjunctive conditionals above with an extension of the standard dynamic treatment of donkey anaphora in DPL (Groenendijk and Stokhof 1991). The resulting system is quite similar to the integrated dynamic framework for anaphora and modality developed in Groenendijk et al. 1996. It differs by defining the meaning of indefinites in terms of several other independently needed operations.\footnote{The framework presented here also differs in sacrificing certain formal features for increased simplicity. Unlike Groenendijk et al. 1996, the present system cannot account for overwrite of variables. For example, for all $\sigma$, $\sigma[\exists x F x]$ is a fixed point of $[\neg \exists x F x]$. This is the cost of eliminating pegs from Groenendijk et al. 1996. The problem can be artificially avoided by stipulating that $[\exists x F x]$ is defined only on those contexts $\sigma$ such that there is no $(g, w) \in \sigma$ for which $g(x)$ is defined.}
To begin with, we define a first order language, and an enriched notion of context. In this framework, assignments are partial functions from variables to individuals, and contexts are sets of assignment-world pairs.

**Definition 6.** Pred$^n$ is the set of $n$-ary predicates $F^n, G^n, \ldots$. Var is the set of variables $x_1, x_2, \ldots$. $\mathcal{L}^+ = \bigcup_{n \in \mathbb{N}} \{F^n(x_1, \ldots, x_n) \mid x_1, \ldots, x_n \in \text{Var} \land F^n \in \text{Pred}^n\}$. $\mathcal{L}^+$ is the closure of $\mathcal{L}_\alpha^+$ under $\exists x_i$ and every operator in $\mathcal{L}$.

$D$ is the domain of individuals. An assignment $g \in G$ is a partial function from variables to individuals. If $g(x)$ is undefined, $g[x]g'$ iff $g'$ extends $g$ by supplying a value for $x$. If $g(x)$ is defined, then $g[x]g'$ iff $g$ and $g'$ differ at most in the value of $x$. Worlds $w, w', \ldots \in W$ are functions from predicates into tuples of individuals. A context $\sigma$ is a set of assignment-world pairs. An extended model $M^+ = (W, [\cdot], f, D)$.

Surprisingly, we can assign almost identical semantic clauses to the fragment without indefinites as in our earlier theory. There is just a single change that we must make to the framework. In the earlier semantics, the question we asked was what it took for a world to be removed from a context during update. Now that contexts contain partial assignment functions, an assignment can be removed during update even when an extension of it remains. In that case, we will say that the assignment still subsists in the resulting context.

Disregarding indefinites, we retain almost the same semantics and notion of entailment as before, replacing membership with subsistence.

**Definition 7.**

1. $(g, w)$ subsists in $\sigma$ $(\langle g, w \rangle \in \sigma)$ iff $\exists h : g \subseteq h \land \langle h, w \rangle \in \sigma$.
2. $\sigma$ subsists in $\sigma'$ $(\sigma \in \sigma')$ iff $\forall(g, w) \in \sigma : \langle g, w \rangle \in \sigma'$.
3. $\sigma$ supports $\phi$ $(\sigma \models \phi)$ iff $\sigma \in \sigma[\phi]$.
4. $\phi_1; \ldots; \phi_n \models \psi$ iff for every $\sigma$, $\sigma[\phi_1] \ldots [\phi_n] \models \psi$.

**Definition 8.**

1. $\sigma[F^n(x_1, \ldots, x_n)] = \{\langle g, w \rangle \in \sigma \mid \langle g(x_1), \ldots, g(x_n) \rangle \in w(F^n)\}$
2. $\sigma[\neg \phi] = \{\langle g, w \rangle \in \sigma \mid \langle g, w \rangle \not\in \sigma[\phi]\}$
3. $\sigma[\Box \phi] = \{\langle g, w \rangle \in \sigma \mid \sigma[\phi] \neq \emptyset\}$
4. $\sigma[\Diamond \phi] = \{\langle g, w \rangle \in \sigma \mid \sigma[\phi] \neq \emptyset\}$
5. $\sigma[\phi \rightarrow \psi] = \{\langle g, w \rangle \in \sigma \mid \sigma[\phi] \models \psi\}$

$F^n(x_1, \ldots, x_n)$ eliminates points from a context where the referent of $x$ is not in the extension of $F^n$. $\neg \phi$ eliminates the points from a context that would subsist in an update with $\phi$. $\Diamond$, $\Box$, and $\rightarrow$ work as before, except $\Box$ and $\rightarrow$ are now defined in terms of our enriched notion of support.

To model subjunctive mood, we introduce assignment-dependent propositions. The proposition that $\phi$, relative to $g$, is the set of worlds $w$ where $\langle w, g \rangle$ supports $\phi$.

**Definition 9.**
1. $[\phi]^{\sigma} = \{w \mid \{g, w\} \models \phi\}$.
2. $\sigma[\bowtie \phi] = (\bigcup \{\{g\} \times f(w, [\phi]^{\sigma}) \mid \langle g, w \rangle \in \sigma\})[\sigma]$.

$\sigma[\bowtie \phi]$ is the result of (i) finding the set of $\langle g, w'\rangle$ where $\langle g, w \rangle$ is in the input and $w'$ is among the nearest $[\phi]^{\sigma}$ worlds to $w$; and (ii) updating that set with $\phi$, to recover any discourse referents introduced by $\phi$.

To interpret indefinites, we need to introduce the main new theoretical idea of this paper: quasi-distributivity. So far, we have enriched contexts to contain referential information as well as worldly information. This has given us a set of assignment-world pairs. But once our contexts contain these two dimensions of information, we must take care regarding how this information is structured. Any given context can contain points that differ with respect to both assignments and worlds. But indefinites and ultimately possibility modals both require this information to be more finely structured. In particular, indefinites require us to partition contexts into cells that agree on the assignment function. To accomplish this task, we introduce a new operator $\#$, which induces a partition on its input context. So $\sigma[\#]$ is a set of contexts ($\Sigma$). In particular, $\sigma[\#]$ partitions $\sigma$ into the subsets of $\sigma$ that share assignment functions, allowing us to consider each potential value of a discourse referent in turn.

Dynamic semantics involves a distinction between updating a context globally and updating every member of that context individually.

**Definition 10.** $\phi$ is distributive iff $\forall \sigma: \sigma[\phi] = \bigcup \{\{g, w\}[\phi] \mid \langle g, w \rangle \in \sigma\}$.

Tests like $\Diamond \phi$ and $\Box \phi$ above are not distributive. But in DPL, indefinites are actually distributive. When integrating a non-distributive modal semantics with a distributive semantics for indefinites, there is then a choice point about whether or not to make updating proceed pointwise. To resolve this choice, we let updating with the indefinite $\exists x \phi$ be quasi-distributive. Exploiting $\#$, indefinites update a context $\sigma$ in a manner that is pointwise with respect to assignments, but global with respect to worlds.

**Definition 11.** $\phi$ is quasi-distributive iff for every $\sigma$:

$\sigma[\phi] = \bigcup \{\{g, w\} \mid g = g' \models [\phi] | g' \in G\}$.

In particular, $\exists x \phi$ will partition $\sigma$ into cells that agree on assignments. Each cell is then updated pointwise, and the results are collected together. So the overall update will be 'cellwise' rather than 'pointwise': it operates globally over a set of worlds and a single assignment, and repeat this process for each assignment in the original context; then the results are unioned together.

To implement this idea, we need to introduce two more tools. First, we must allow indefinites to extend the assignment functions that are already available in a context. To accomplish this task, we introduce an operation $[\exists x]$ that essentially adds a new discourse referent for $x$. In particular, $\sigma[\exists x]$ is the set of $\langle g, w \rangle$
such that $g$ is a minimal $x$-variant of some $g'$ and $(g', w) \in \sigma$. In general, this will involve taking a $g'$ which is not defined for $x$, and extending it with some value for $x$. So updating with the indefinite phrase $\exists x \phi$ involves first introducing a new discourse referent with $\exists x$, and next partitioning the resulting context with $\#$. To finish the update, we introduce a final tool, $\downarrow$, which lets us merge this set of contexts into a single context. In particular, where $\Sigma$ is a set of contexts $\sigma$, $\Sigma[\downarrow \phi]$ is the result of applying $\phi$ to each element of $\Sigma$ and taking the union of the result. With these tools in hand, we can give a semantics for the larger construction $\exists x \phi$.

**Definition 12.**

1. $\sigma[\exists x] = \{(g', w) \mid \exists (g, w) \in \sigma : g(x) = g'\}$
2. $\sigma[\#] = \{(g, w) \in \sigma \mid g = g' \mid g' \in G\}$
3. $\Sigma[\downarrow \phi] = \bigcup\{[\sigma[\phi] \mid \sigma \in \Sigma\}$
4. $\sigma[\exists x \phi] = \sigma[\exists x][\#][\downarrow \phi]$

Summarizing, we factorize the effect of $\exists x \phi$ into four successive operations. First, $\exists x \phi$ introduces a discourse referent for the variable $x$, by allowing $x$ to take any value whatsoever. $\exists x \phi$ next partitions the resulting context into cells which agree on their assignment. Third, it updates each cell with $[\phi]$. Finally, it takes the union of the result. While this process is complex, it is also well behaved: when $\phi$ is distributive, $\sigma[\exists x][\#][\downarrow \phi] = \sigma[\exists x][\phi]$. That is, the purpose of quasi-distributivity is to assist in the interaction between indefinites and modals.

We are now in a position to see how this framework accommodates the essential features of donkey conditionals while ascribing (6)-(8) the LF in (9).

$$\exists x \phi \rightarrow \psi \tag{9}$$

The first feature of donkey conditionals was that the denotation of the pronoun in the consequence co-varies with the antecedent’s indefinite, despite occurring outside of that indefinite’s scope. This occurs because an existential quantifier has an effect on context which is not limited to its syntactic scope. Once $\sigma$ is updated with $\exists x \phi$, future uses of $x$ will refer to an individual that is $F$. So the quantifier in the antecedent of (9) can shift the assignments at which variables in the consequent are evaluated, despite not binding them directly.

The second important feature of donkey conditionals we discussed above was universal interpretation. The semantics above explains this through the interaction of the meaning of indefinites and the conditional. The conditional $\phi \rightarrow \psi$ requires that $\psi$ is supported after updating with $\phi$. This in turn requires that $\sigma[\phi]$ subist in $\sigma[\phi][\psi]$, so that every $(g, w)$ pair in $\sigma[\phi]$ has some extension in $\sigma[\phi][\psi]$. When $\phi$ is of the form $\exists x \phi$, $\sigma[\phi]$ will contain an assignment mapping $x$ to each value of $F$. Subsistence thus requires every $F$ in the domain to satisfy the consequent. This gives universal force to existential quantifiers in the antecedent of conditionals.
5 Subjunctive Donkey Conditionals

Section 4 integrated a dynamic theory of subjunctive conditionals with a dy-
namic theory of anaphora in order to accommodate the behavior of subjunctive
donkey conditionals. Our first task here is to explain the selective / unselective
ambiguity, regarding which individuals are quantified over by donkey condition-
als. We explain this in terms of the relative scope of indefinites and subjunctive
mood.

In particular, our theory allows two different representations of subjunc-
tive donkey conditionals like (1) and (2). Either \( \langle \cdot \rangle \) can outscope \( \exists x \), or \( \exists x \) can
outscope \( \langle \cdot \rangle \).

\[
\begin{align*}
(10) & \quad \langle \exists x Fx \rangle \to \phi \\
(11) & \quad \exists x \langle Fx \rangle \to \phi
\end{align*}
\]

In our theory, (10) and (11) correspond to selective and unselective readings of
the conditional, respectively. Consider their effect in turn.

[(10)] tests \( \sigma \) to check whether \( \sigma[\langle \exists x Fx \rangle] \) supports \( \phi \). This in turn requires
that \( \sigma[\langle \exists x Fx \rangle] \) subsists in \( \sigma[\langle \exists x Fx \rangle][\phi] \), which requires that every \( (g, w) \in \sigma[\langle \exists x Fx \rangle] \) has an extension in \( \sigma[\langle \exists x Fx \rangle][\phi] \). \( \sigma[\langle \exists x Fx \rangle] \) is the set containing, for
each \( (g, w) \in \sigma \), every \( (g', w') \) such that \( w' \) is among the nearest \( \llbracket Fx \rrbracket^g \)-worlds to \( w \), and \( g'(x) \in w'(F) \). Crucially, there can be objects in the domain which are
only \( F \) at worlds far removed from the context. These objects are not considered
discourse referents for \( x \) by \( \langle \exists x Fx \rangle \), because \( f \) simply selects the closest worlds
where there is some \( F \). So this reading only quantifies over \( D_{near} \), the set of
objects that are \( F \) at the nearest worlds where something is \( F \). \( \sigma \) supports (10)
iff all such objects satisfy \( \phi \).

[(11)] is also a test. \( \sigma \) supports [(11)] iff \( \sigma[\exists x \langle Fx \rangle] \) supports \( \phi \). This in turn
requires that \( \sigma[\exists x \langle Fx \rangle] \) subsists in \( \sigma[\exists x \langle Fx \rangle][\phi] \), which requires that every \( (g, w) \in \sigma[\exists x \langle Fx \rangle] \) has an extension \( (g', w) \) in \( \sigma[\exists x \langle Fx \rangle][\phi] \). \( \sigma[\exists x \langle Fx \rangle] \) is the set
\( \sigma[\exists x][\#]\langle \cdot \rangle \). Since \( \langle \cdot \rangle \) is distributive, this is equivalent to \( \sigma[\exists x][\langle \cdot \rangle \phi] \). Thus,
\( \sigma[\exists x \langle Fx \rangle] \) is the result applying \( [\cdot Fx] \) to the set of \( (g', w) \) such that \( (g, w) \in \sigma \)
and \( g[x]g' \). Thus it returns, for each \( (g', w) \) in this set, the unique \( (g', w') \) such
that \( w' \) is the nearest \( \llbracket Fx \rrbracket^g \)-world to \( w \). Crucially, for each \( d \in D \), there is
some \( (g', w') \in \sigma[\exists x] \) such that \( g'(x) = d \). So we take every individual \( d \) in the
domain, construct a \( g \)-alternative that assigns \( x \) to \( d \), and then find the closest
worlds to those in the context where \( d \) is \( F \). \( \sigma \) supports (11) iff the resulting
set of world-assignment pairs supports \( \phi \). Here, the key idea is the fundamental
distributivity of indefinites with respect to assignments. For each individual, we
must find the closest worlds where they are \( F \). This is because \( \exists x \) introduces
an assignment for each individual, and \( \langle Fx \rangle \) distributively moves through each
assignment-world pair, finding the closest worlds for each value of \( x \). Similarly,
in the next section we will also explain the existential / universal ambiguity by
attending to the question of how much distributivity different updates require.

Summing up, \( \sigma[\langle \exists x Fx \rangle] \) contains assignments mapping \( x \) to whoever is \( F \)
at the closest worlds to the context where something is \( F \), paired with those
worlds. By contrast, $\sigma[\exists x \triangleleft Fx]$ contains an assignment for every relevant individual, paired with the closest worlds to the context where that individual is $F$. When embedded in an antecedent, the former gives rise to a selective reading of the conditional, whereas the latter gives rise to an unselective reading. We have derived our two readings using the simple assumption that in subjunctive conditionals with indefinite antecedents, the antecedent can take either wide or narrow scope relative to mood. Furthermore, since nothing depends upon the content of the consequent, we account for the selective / unselective reading in full generality across both might and would conditionals.

6 Might Conditionals

Our explanation of the selective / unselective ambiguity in terms of relative scope is a new theoretical result. But it doesn’t provide a significant empirical advantage for our analysis compared to earlier theories. For this, we now turn to the existential / universal ambiguity. Again, the question here is whether donkey might counterfactuals require that every relevant satisfier of the antecedent satisfies the consequent, or merely some of them. The framework above can explain this by relying on the quasi-distributive partition operator $\#$. They key idea is that possibility modals can exploit $\#$ to update a context cell-wise rather than the context globally. This more complex operation removes any assignment whose choice of referent cannot satisfy the prejacent.

It turns out that the existential reading of might conditionals basically comes for free on the analysis above. Under Definition 8, $\sigma \models \phi \rightarrow \Diamond Fx$ iff there is some $(g, w) \in \sigma[\phi]$ such that $g(x) \in w(F)$. This corresponds to the existential reading of the conditional. Crucially, this is because $\Diamond$ is an essentially existential test: it checks that the entire context can be updated with $Fx$ consistently. This allows that $Fx$ may remove some potential values of $x$.

In order to generate the universal reading of the conditional, we require a test which passes at $\sigma$ iff for each $g$ such that $(g, w) \in \sigma[\phi]$, there is some $(g, w') \in \sigma[\phi]$ such that $g(x) \in w'(F)$. That is, we need an operation that considers every assignment function separately, requiring that this assignment function maps $x$ to an individual that is $F$ at some world in the context. It turns out that this can be easily obtained by applying the $\#$ and $\downarrow$ operators to the consequent. In particular, a natural idea is to use $\#$ to first partition the context by assignments, and then use $[\downarrow \Diamond \phi]$ to test each cell of the context to check whether it can be consistently updated with $\phi$.

One way to implement this idea is to make might lexically ambiguous. Perhaps, that is, there is a second potential meaning for might ($\Diamond \# \phi$) that directly incorporates quasi-distributivity.

\textbf{Definition 13.} $\sigma[\Diamond \# \phi] = \sigma[\#][\downarrow \Diamond \phi]$

\footnote{In fact, something like this meaning for might is defended for independent reasons in Büring 1998, where it is formulated without regards to $\#$ and $\downarrow$.}
Since $\Diamond\phi$ is not distributive, $\sigma[\Diamond\phi] \neq \sigma[\Diamond\#\phi]$. The latter meaning is not a test, because it narrows down a context by filtering out any assignments where the relevant individual cannot satisfy the information contributed by $\phi$. Again, this is because $\Diamond\#\phi$ is quasi-distributive. It treats assignments distributively, and worlds globally.

We now treat the existential and universal readings of the conditional as follows:

(12) $\phi \rightarrow \Diamond Fx$

(13) $\phi \rightarrow \Diamond\# Fx$

[[12]] tests $\sigma$ to see whether $\sigma[\phi]$ supports $\Diamond Fx$. This in turn requires that every $\langle g, w \rangle$ in $\sigma[\phi]$ subsists in $\sigma[\phi][\Diamond Fx]$. Since $\Diamond Fx$ is a test, this last condition is equivalent to the requirement that $\sigma[\phi]$ can be consistently updated with $Fx$. This in turn requires merely that some assignment-world pair in $\sigma[\phi]$ maps $x$ to an individual that is $F$ in the relevant world. This is an existential truth condition.

[[13]] tests $\sigma$ to see whether $\sigma[\phi]$ supports $\Diamond\# Fx$, and thus whether every $\langle g, w \rangle$ in $\sigma[\phi]$ subsists in $\sigma[\phi][\Diamond\# Fx]$. This in turn requires taking each assignment $g$ in $\sigma[\phi]$ and pairing it with each world $w$ where $\langle g, w \rangle$ is in $\sigma[\phi]$. For $\langle [13] \rangle$ to be supported, every such cell of $\sigma[\phi]$ must be consistent with $Fx$. That is, every $g$ in $\sigma[\phi]$ must map $x$ to an individual that is $F$ at some world in $\sigma[\phi]$. So $\langle [13] \rangle$ ultimately requires that every satisfier of the antecedent is possibly $F$.

Summing up, a context supports (12) iff there is some pair of an assignment $g$ and world $w$ where $g$ maps $x$ to an individual that satisfies the antecedent, $w$ is one of the closest worlds to the context where $x$ satisfies the antecedent, and $x$ is $F$ at $w$. By contrast, a context supports (13) iff every assignment-world pair that survives updating with the antecedent maps $x$ to an individual that is $F$ at some world in the context.

We now have an analysis of both the selective / unselective ambiguity, and the existential / universal ambiguity. Combining our two analyses, (14a)-(14d) will generate readings (1a)-(1d) respectively.

(14) \begin{align*}
\text{a. } & \exists x < Fx \rightarrow \Diamond\# Gx & \text{UNSELECTIVE UNIVERSAL} \\
\text{b. } & <\exists x Fx \rightarrow \Diamond\# Gx & \text{SELECTIVE UNIVERSAL} \\
\text{c. } & <\exists x Fx \rightarrow \Diamond Gx & \text{SELECTIVE EXISTENTIAL} \\
\text{d. } & \exists x < Fx \rightarrow \Diamond Gx & \text{UNSELECTIVE EXISTENTIAL}
\end{align*}

To conclude, we note this analysis elegantly explains why the existential / universal ambiguity is absent for would conditionals. Bare would conditionals do not contain a modal in the consequent. In general, quasi-distributivity is a property that only effects modal sentences: $\#$ has no important effect when it applies to an ordinary update. For ordinary updates are fully distributive, and hence it does not matter whether they apply cell by cell, or all at once. More precisely, for any non-modal $\phi$, we have that $\cup(\sigma[\#][\phi]) = \sigma[\phi]$. So it doesn’t matter whether $\phi$ updates each cell of a partition, or whether $\phi$ updates the union of that partition. The same reading is achieved.
Bibliography


1 Introduction

The current study develops a substantive link between three independent and chronologically separate diachronic trajectories in Classical Chinese syntax. All three trajectories involve a similar pattern of development: a newly copularized (reanalyzed from lexical sources) morpheme emerged in both copular clauses and clefts, and subsequently its use in both constructions have declined. The recurring trend suggests a non-coincidental connection between the copula morphemes in both copular clauses and clefts that point to syntactic and semantic relatedness.

Specifically, the historical evidence lends support to the claim that clefts in Chinese are a type of copular construction, following some steps of transformation. This homogeneity approach contrasts with an alternative approach that treats the copula morpheme in clefts as participating in focus movement together with the focused phrase. In the current study, I show that the latter approach faces difficulties in addressing the coordinated emergence and decline pattern in my diachronic data. Moreover, when it comes to semantics, a derivation of clefts that assumes an underlyingly copular structure is readily available, leaving it less parsimonious to adhere to a focus movement story.

The rest of this paper is structured as follows: section 2 briefly presents three copula and cleft trajectories for the copula morphemes \textit{wei}, \textit{shi}, and \textit{xi}, respectively. Section 3 discusses how a comparative analysis of the three distinct trajectories gives credence to the underlying semantic uniformity between copular clauses and clefts in Chinese. Section 4 concludes the paper.

2 Data

The three grammatical change trajectories of Classical Chinese copula morphemes \textit{wei}, \textit{shi}, and \textit{xi} are shown in Table 1.

Examples (1-3) exemplify these diachronic developments. (a)-examples illustrate copular clause uses, and (b)-examples demonstrate cleft uses. In the copular clause examples, VP-modifiers (e.g. the future tense operator/modal auxiliary


<table>
<thead>
<tr>
<th>Copula morpheme</th>
<th>Grammatical change pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wei</td>
<td>Emergence (reanalysis) of copula &amp; cleft (1300<del>1050 BC) → (Copula &amp; cleft) Simultaneous decline in Classical Chinese (600</del>100 BC)</td>
</tr>
<tr>
<td>Shi</td>
<td>Emergence (reanalysis) of copula &amp; cleft (~100 AD) → (Copula &amp; cleft) Simultaneous decline in modern Cantonese (by 18th century AD)</td>
</tr>
<tr>
<td>Xi</td>
<td>Emergence (reanalysis) of copula &amp; cleft (~900 AD) → (Copula &amp; cleft) decline in modern Mandarin (by 18th century AD)</td>
</tr>
</tbody>
</table>

Table 1. Summary of the three pathways

qi, the adverb *yuan* ‘originally’ and the negation word) are shown to precede copulas, which demonstrates the latter’s verbal status.1

(1) a. ya bín qǐ wèi chén.
   noble Bin FUT/MOD WEI minister
   “The noble Bin will be minister.”
   (*Collections of Oracle Bones*, 22301, 13th century -11th century BC)

b. wèi [dǐ] tā wǒ nián.
   WEI god place.a.curse my yearly.harvest
   “It is god who put a curse on my crops.”
   (*Collections of Oracle Bones*, 6476, 13th century -11th century BC)

(2) a. Yizhou yuan bu shì yongrén.
   Yizhou originally NEG SHI mediocre.person
   “Yizhou was originally not a mediocre person.”
   (*Lectures on the Vimalakirti Sutra*, 8th century-9th century AD)

b. yìqi zhōngshēng píngdēng, jie you zhēnrufáshēn, fēi shì
   all lifes be.equal, DIST have dharma, NEG SHI
   [you du] shinai you ye.
   through deliverance start have DECL.PRT
   “All lifes are equal, and all are capable of obtaining dharma. It is
   not via deliverance that we start to obtain (dharma).”
   (*Lectures on the Vimalakirti Sutra*, 8th century-9th century AD)

(3) a. Líng zhēnshòu neíchen yuan yì bù xì
   make garrison.officer interior.officer originally also NEG XI
   hóngwù jiuzhí.
   First.Emperor old.law
   “Making the garrison officers part of the interior ministry was originally also not an old law made during the reign of the First Em-

---

1 Word order correlations have remained stable across different historical periods [24, 20, 26, 17]. The most common word order is verb-medial for transitive sentences. In terms of phrase-internal constituents, the modifier tends to precede the modified (e.g. VP modifier < verbal head). Finally, despite the predominant SVO word order, Chinese is consistently topic-prominent, allowing object and oblique preposing [23].

85
peror."

(Collections of Yu Qian, 9. 57, 1398 - 1457 AD)

b. Xi [shou ningxia zongbiingguan] chuzhi,  shili wei gan Xi by Ningxia head.offer deal.with, affairs NEG dare.to
determine
“It is by the head officer of Ningxia that the affairs are dealt with, therefore regarding the affairs, I did not dare to make my own call.”

(Collections of Yu Qian, 1.10, 1398-1457 AD)

All the three copula morphemes underwent independent decline processes: *wei* disappeared from copular clauses and clefts by the 6 - 4th century BC. *Shi* remains in copular clauses and clefts in modern Mandarin, yet has ceased to be used in the two constructions in other Sinitic languages such as Cantonese (all Sinitic languages are descendants of Classical Chinese). *Xi* exhibits the opposite pattern: productivity in copular clauses/clefts is retained in modern Cantonese but not in Mandarin [27].

Specifically, it is argued that *wei* was reanalyzed into a focus-sensitive, only-like adverbial operator. Although a quantitative study of this process is not available due to the scarcity of corpus sources at this period, the reanalysis of *wei* can be evidenced in a series of changes in its syntactic behaviors. First, *wei* became optional in the cleft construction [23]. In addition, it is witnessed to co-occur with the copula *shi* [26, 23]. Furthermore, *wei* was able to associate with any focus in its scope, including focused constituents not immediately adjacent to *wei* [30]. These behavioral changes are compatible with a focus-sensitive adverbial analysis. Finally, the order of *wei* preceding *bu* ‘negation’ is also attested (while a copula *wei* must follow a modifying negative morpheme), which receives an explanation if both morphemes are adverbs of quantification (so that their relative order is flexible and transparently reflects their scopal relations) [26, 30].

In modern Mandarin, the decline of *xi*’s use in copular clauses and clefts is manifested in the token frequency of *xi* vis-à-vis the copula *shi* in spoken Mandarin corpora. As *shi* is additionally used as a frequent component morpheme in a variety of high-frequency compound words (e.g., *dan-shi* ‘however’, *yu-shi* ‘then’), this study restricts the token occurrences of *xi* versus *shi* to copulative sentences of the form [*Ta shi/xi X.*] ‘(S)he is X (end with a period punctuation mark)’. I first investigate the electronic database of *People’s Daily*, a broadsheet newspaper employing a formal style. The results reveal that a total of 1349 *xi*-copulas (631 occur in copular clauses that end with an NP constituent; 718 in cleft clauses) are used against a total of 1901 *shi*-copulas (944 copular: 957 cleft). In stark contrast to the findings in *People’s Daily*, the distribution of the *xi*-copula versus the *shi*-copula is clearly asymmetric in the colloquial *Duzhe* corpus (a popular family magazine): against an aggregate 24 *xi*-copula tokens in the years 1996-2012 (ratio of copular clause/cleft clause: 17:7), an average of 3513 tokens of the *shi*-copula are found per year (ratio of copular clause to cleft clause 58:17). In sum, corpus studies suggest that Mandarin *xi*-copula has all but disappeared from colloquial discourse.
Unlike modern Mandarin, in spoken Cantonese corpora, the \textit{xi}-copula is prevalent in colloquial use, whereas \textit{shi}-copula is exceedingly rare. In a Hong Kong Cantonese corpus (HKCanCor), which collects 93 radio talk shows and telephone transcripts from 1997 to 1998, only two \textit{shi}-copula tokens, compared with a total of 5387 \textit{xi}-copulas. In opposition, a total of 5387 \textit{xi}-copulas are found in copular clauses and clefts. This asymmetry thus establishes the \textit{xi}-copula’s prevalence in modern Cantonese. In short, it seems that modern daughter languages of Classical Chinese opt for one single copula morpheme in fulfilling the role in copular clauses and clefts, while the other copula morpheme inherited from the ancestor language tends to be demoted from both constructions.\footnote{Corpus studies on the development/decline of copula morphemes are not available for the other Sinitic languages. What we know is that for almost all modern Sinitic languages, if a given Sinitic language uses \textit{shi}-copular clauses, it also uses \textit{shi}-clefts, and not \textit{xi}-clefts. If a given Sinitic language uses \textit{xi}-copular clauses, it also uses \textit{xi}-clefts instead of \textit{shi}-clefts (e.g. Li and Thompson 19, 20; Djamouri 4).}

### 3 Towards a copular approach to Chinese clefts

In the following, I present an argument that the recurring diachronic trajectory of the coordinated development of copular clauses and clefts lends support to analyzing clefts as copular clauses. I start by reviewing two alternative approaches to Chinese clefts and then explain how the copular approach is favored by the diachronic data.

#### 3.1 The copular and the focus movement approach: A summary

The copular approach to Chinese clefts posits that a cleft is underlingly a copular structure, headed by the copula as the matrix predicate. One analysis, championed by Li and Thompson [20], characterizes Chinese clefts as inverted pseudoclefts. The subject of the copula verb is occupied by a relative (with a null operator head) taking a covert definiteness determiner. In surface syntax, the relative extraposes to the right of the focused phrase.\footnote{This treatment of clefts as concealed pseudoclefts resembles proposals for English \textit{it}-clefts \cite{1,25}, barring that there is no overt \textit{it}-subject in Chinese.} An implementation of (4-a) in line with this analysis is provided in (4-b).\footnote{Several proposals, e.g. Zhan and Traugott 31, contend that the combination of a definiteness operator and a relative head is evidenced by the optional surface realization of the particle \textit{de} that attaches to the relative clause in Chinese. However, this analysis of clausal particle \textit{de} remains controversial \cite{12}. This paper does not discuss further details for space reasons. My primary concern is to motivate the general theoretical position that clefts are underlingly copular structures, rather than the merits of a specific treatment.}

\begin{exe}
\ex \label{4a} Shi Zhangsan yao lai.
   SHI Zhangsan will come
   “It is Zhangsan that will come.”
\end{exe}
b. [DEF e_i] shi [Zhangsan] [Op [yao lai]],

Another proposal, articulated by Hole [12], claims that the copula verb takes a small clause (CP) argument. The focused phrase moves to [Spec, CP] of the small clause, and predicate abstraction applies to the CP predicate. The C^0 head functions as a special definite determiner, deriving a maximal event reading of the CP predicate (in comparison with a regular definite determiner that takes in a nominal predicate and yields a maximal individual).

(5) [VP shi [CP [Spec Zhangsan]i [C’ C^0[+def] [TP e_i [VP yao lai]]]]]

The focus movement approach adopts a monoclausal analysis, in which a Chinese cleft is not headed by a copula verb. Rather, the copula moves with the focused phrase to the left periphery to check the [exhaustive] feature (Teng 28, Huang 14, Zhu 32). Assuming a Rizzi-style articulated CP, it is argued that the focused phrase undergoes focus movement to [Spec, FocP] from its base position at FinP. One characterization of the copula morpheme’s role during focus movement is that shi is syntactically an adverb analogous to the English adverb only [14, 32, 5]. An adverb-like focus marker, pace Rooth, resides in the left periphery but simultaneously stays as closely to the focus it associates with as possible. Another possibility is that the copula morpheme initially merges at the head of the focus projection (taken to be FP by Kiss 15, which is similar to Rizzi’s FocP) and subsequently undergoes remnant movement to a projection structurally higher than FocP (e.g. TopP, cf. Kiss 15, Meinunger 22, Frascarelli and Ramaglia 6). An implementation of (4) under this proposal is in (6). In either analysis, the feature-checking position or landing site of shi is not motivated in a very precise way, except that the position must be above FocP to capture the correct word order.

(6) [TopP shi_k [FocP Zhangsan]i [Foc’ Foc^0 [+e_k [FinP e_i yao lai]]]]

3.2 Evidence in support of the copular approach

I argue that the recurring trajectory identified in historical Chinese receives a straightforward explanation under a copular approach to clefts. Specifically, after reanalysis of a lexical item into a copula takes place, learners acquire the newly copularized item as an element of copula verbs within her/his lexicon. Assuming that lexical insertion is triggered, such that the new element is inserted to the copula verb head position in the syntax, we would expect that the same copula element occurs in all constructions that host a copula verb projection. This includes the cleft construction, which is a copular clause construction within learners’ grammar, according to the copular approach. In other words, the recurring diachronic trajectory is reduced to a reanalysis-and-extension process [8]: the reanalysis of a morpheme as instantiating a copula verb category results in the extension of this morpheme to structures that host the copula verb category. The copular approach also readily accounts for the coordinated decline pattern witnessed in the three morphemes wei, shi and xi: it follows from the
homogeneity of copular and cleft structures that the loss of productivity of a
given copula verb predicts that it will cease to be used in both copular clauses
and clefts.

Here an important distinction needs to be drawn between a focus-based, mon-
oclausal analysis of cleft structures and an analysis arguing that a biclaual cleft
structure has subsequently undergone a further reanalysis process and developed
into a simplex clause featuring a focus adverb. The latter has been argued to be
crosslinguistically common (e.g. Heine and Reh [9]), and for a natural reason:
Since the copula is always adjacent to the focused constituent in cleft struc-
tures, a reanalysis of the copula as a focus marker is often possible. Therefore,
the possibility needs to be left open that copulas in cleft structures may develop
into focus markers, accompanied by a syntactic reconfiguration of a biclaual
cleft structure into a monoclausal structure. Indeed, as Section 2 shows, this is
exactly what happens to the Chinese copula wei, which first appears in clefts,
and subsequently occurs as an adverb in a simplex sentence (both its cleft use
and its copular clause use gave way to an adverbal use).\(^5\)

While a copula-based analysis yields a straightforward and simple explana-
tion of the recurring copula-and-cleft pathway identified in this paper, a focus-
based approach is burdened with finding clear motivations for why copula mor-
phemes demonstrate a recurring trend of moving to a projection within or higher
than FocP (beyond the need to derive the correct word order).

A more severe difficulty is that the focus-based approach fails to account
for the simultaneous decline of the copular and cleft use. As previously shown,
the use of wei in copular clauses and clefts have both disappeared by Classical
Chinese. Furthermore, the use of xi in copular clauses and clefts have also
vanished side by side in modern Mandarin, and in parallel the use of shi in both
constructions have vanished in modern Cantonese. A focus-based analysis of
Chinese clefts would commit to positing two homophonous lexical entries for the
copula morpheme that occurs in the copular clause construction and in the cleft
construction, respectively. A direct consequence is the lack of convincing reason
to account for why the loss of both lexical entries should be closely correlated.\(^6\)

\(^5\) Although a strong crosslinguistic tendency for copulas to appear in cleft construc-
tions has been identified in the previous grammaticalization literature [9], in many
languages it has been argued that clefts are distinct from copular clauses and instanti-
tate a monoclausal structure. In particular, such monoclausal analyses have been
proposed in several languages/language families that are areally close to or typo-
logically similar to Chinese, such as Japanese, Uzbekh and Wolof (cf. Hiraiwa and
Ishihara 10, Gribanova 7, Klecha and Martinović 16), in which cases the cleft mor-
pheme is homophonous with the copula verb. Consequently, I believe it is premature
to treat the biclausal, copular analysis of Chinese clefts as the default analysis. It is
still desirable to look into Chinese-internal evidence when we compare the biclausal
and the monoclausal analysis.

\(^6\) Frascarelli and Ramaglia [6], differing from other focus approaches, makes critical
use of the copula position as the merge site and makes the building up of the deriva-
tion tree hinge upon a copular structure. As a result, what is to some extent a
biclausal copular structure becomes reconfigured (i.e. the copula morpheme merges
3.3 Semantic evidence

One of the central motivations for a focus-movement approach to clefts lies in generating the exhaustiveness reading via movement of the focused phrase to FocP and checking \([\text{exhaustive}]\) feature. This paper will not turn to the implementation details of such approach. Importantly, an exhaustiveness reading can also be derived by assuming a copular structure syntax. In this subsection, I discuss two prominent recent proposals along this line. In both proposals, a dedicated focus projection in the left periphery is not required as the locus of exhaustivity. As a point of departure, these two semantic treatments both assume that the exhaustiveness reading belongs to presupposition, rather than the level of at-issue semantic content.\(^7\) This distinction between two layers of meaning is motivated by the observation, first articulated by Horn [13], that the exhaustive interpretation of clefts differs from the \textit{only}-interpretation. Horn’s original examples are about English \textit{it}-clefts, but the contrast he showed to exist between \textit{it}-clefts and \textit{only}-sentences also holds in Chinese, as in (7).

\begin{itemize}
  \item[(7)] a. *Zhangsan zhidao Lisi yaoqing le Wangwu. Danshi ta bu zhidao shi Lisi yaoqing le Wangwu. But he NEG know Lisi invite PRF Wangwu.
  \item \hspace{1cm} *Zhangsan knows that Lisi invited Wangwu. But he doesn’t know that it is Lisi that invited Wangwu.*
\end{itemize}

\(^7\) The claim that the locus of cleft exhaustivity is at the level of presupposition is widely adopted in more recent semantic proposals. In principle, this position is compatible with both a semantics couched in a copular-structure syntax and a focus-based syntax. See Erlewine [5], for a focus-based analysis where exhaustivity is presupposed.
b. Zhangsan zhidao Lisi yaoqing le Wangwu. Danshi ta bu
  zhidao zhiyou Lisi yaoqing le Wangwu.
  “Zhangsan knows that Lisi invited Wangwu. But he doesn’t know
  that only Lisi invited Wangwu.”

This phenomenon would be hard to explain if the focus interpretation of clefts
is equivalent to only-sentences. In contrast, if we assume that the exhaustive
reading shi-clefts generates is not part of what is asserted, the oddness in (7-a)
is accounted for: at the at-issue level, the second sentence negates what is asserted
by the first sentence.

Building on Horn’s presuppositional account, Büring and Križ [2] propose
that the interpretation of clefts is computed based on an inverted pseudocleft
structure in the spirit of Percus [25], where the extraposed, nominalized cleft
clause bears a covert definite determiner. The crucial piece of argument is that
the definite description projects a presupposition, presented as follows:

(8) A structure of the form [shi a P]
   Asserts: [[P]][[[a]]]
   Presupposes: [[a]] is not a proper part of [[P]].

Applying (8) to derive the exhaustive reading in (4) yields the following.
According to a Percus-style pseudocleft syntax, the cleft clause takes a covert
definite operator (we abstract away from the extraposition operation here). As
a definite description, it projects the presupposition that the referent of the cleft
phrase Zhangsan is not a proper part of the referent of the cleft clause, yao lai
‘will come’:⁸

(9) Shi Zhangsan yao lai.
   SHI Zhangsan will come
   Asserts: [[will come]][[[Zhangsan]]]
   Presupposes: [[Zhangsan]] is not a proper part of [[will come]].

Consider a situation in which more than one individual will come (e.g.
Zhangsan and Lisi). The atomic individual denoted by [[Zhangsan]] is a proper
part of [[will come]], falsifying the presupposition. Secondly, if no individuals
will come according to the situation, [[Zhangsan]] is not a proper part of [[will
come]] and the presupposition is satisfied. However, the at-issue semantic con-
tent (i.e. what is asserted) will be false. Thus, the only way to satisfy both the

---

⁸ Intuitively speaking, a proper part of something is simultaneously a part of that
thing but not equal to that thing. In set-theoretic terms, each of {a}, {b} and {a,b}
is a part of {a,b}. However, only {a} and {b} are proper parts of {a,b}. {a,b} is not
a proper part of itself.
presupposition and the at-issue content is for Zhangsan to be the maximal (only) individual who will come, hence an exhaustive reading.\footnote{With the assumption that a definiteness operator combines with the cleft clause, a cleft sentence such as (9) is paraphrased as: ‘The one that will come is Zhangsan.’ If we further assume with most current theories of definiteness (e.g. Heim) by positing that the definite operator yields a maximal individual, then the above paraphrase is equivalent to saying that the maximal individual that will come is Zhangsan. A careful reader might then wonder whether it would suffice to derive the exhaustiveness reading solely by the maximality operator. Importantly, the limits of the maximality operator lie in negative contexts. Consider a cleft sentence it’s not Zhangsan that will come. #Lisi will also come. The problem with this utterance stems from the maintenance of exhaustiveness under negation: the cleft sentence is true if and only if one individual alone will come (albeit not Zhangsan). The definite maximality operator won’t predict this. Assuming that both Zhangsan and Lisi will come. The paraphrase of the cleft sentence would be: ‘the maximal individual that will come is not Zhangsan.’ Under this paraphrase, the sentence should be true, contrary to fact. Note that a presuppositional account is able to explain the exhaustiveness under negation, since Zhangsan is a proper part of the maximal individual that will come and hence the presupposition is violated. See [2] for detailed discussions.}

It is worth noting that Büring & Križ’ treatment does not crucially hinge upon a pseudocleft structure. In subsection 3.1, we have seen that, under a Hole-style syntax, the cleft clause within a small clause undergoes no nominalization, but it similarly receives a definite reference, by denoting a maximal event [12]. We can thus see that the exhaustive reading derived from definite descriptions is also compatible with Hole’s implementation.

In another proposal, Velleman et al. [29] posit a covert propositional operator, termed CLEFT, that attaches to the sentence radical in a cleft sentence. CLEFT takes the sentential argument’s proposition as input and outputs a two-layered meaning: the first layer is a minimal (weak) meaning, which states that the proposition \( p \) that the operator combines with has a true answer that is at least as strong as \( p \). The second layer is a maximal (strong) meaning, which states that there exists no strictly stronger true answer than \( p \). Velleman et al. argue that CLEFT projects the strong meaning as a presupposition and asserts the weak meaning. Specifically, Velleman et al. argue that the semantics of the CLEFT operator is defined as follows [29] (the subscript \( S \) indexes the current context):

\[
\text{CLEFT} = \lambda w. \lambda p: \text{MAX}_{S}(p)(w). \text{MIN}_{S}(p)(w)
\]

Where
\[
\text{MIN}_{S}(p)(w): \text{There is a true answer at least as strong as } p \text{ at } w.
\]
\[
\text{MAX}_{S}(p)(w): \text{No true answer is strictly stronger than } p \text{ at } w.
\]

When it comes to the exhaustive reading of clefts, Velleman et al. [29] predict the same desired result as Büring and Križ [2]. Consider the case where CLEFT combines with the sentential argument it is Zhangsan that will come. As per (10), the sentential argument of CLEFT feeds the corresponding proposition \( p: \text{“Zhangsan will come”} \) as the input to CLEFT. The at-issue semantic content guarantees that what is asserted is at least as strong as the proposition Zhangsan
will come. Importantly, the cleft sentence also presupposes that strictly stronger assertions than Zhangsan will come are ruled out. In cases where Zhangsan will not come, the minimal assertion is not satisfied. On the other hand, in cases where Zhangsan is not the only person who will come, the presupposition is violated, since a strictly stronger answer is available. Velleman et al. are less explicit about how surface structure maps to the semantic interpretation. Crucially, nevertheless, the covert CLEFT operator in their account takes propositional arguments as input. As such, this treatment is compatible with a semantic solution that is devoid of focus movement of the cleft phrase.

Since both Büring and Križ’ and Velleman et al.’s proposal can be minimally adapted to fit into an underlyingly copular syntactic structure, it is not necessary to commit ourselves to a particular semantic implementation. It suffices to note that these two semantic proposals are capable of capturing the focus interpretation without positing movement to a left periphery projection. All else being equal, a non-movement semantic implementation is favored by the diachronic pattern of Chinese copula morphemes.\(^\text{10}\)

### 3.4 Structural evidence

Syntactically, focus movement predicts island sensitivity. As Hiraiwa and Ishihara 10 and Hiraiwa and Ishihara [11] observe, ungrammaticality arises when the cleft clause in Japanese da-clefts occur in a strong island environment (such as complex NP islands. This is taken as evidence that the focused constituent undergoes syntactic movement in clefts. It would seem that under a similar structural analysis of Chinese shi-clefts, the cleft phrase also moves to FocP and hence is predicted to induce island effects. However, this prediction is not borne out. The following constructed complex NP island sentence causes no ungrammaticality (based on three Northern Mandarin speakers I consulted):

\[(11) \quad \text{Shi [napian lunwen]}_{i} \text{ ta xiangxin hui you [ken jieshou e] de pingwei.} \]

\[\text{SHI [that paper]}_{i} \text{ he believe will have willing.to accept e REL reviewer} \]

“\text{It is [that paper]}_{i} \text{ that he believes there will be reviewers [who are willing to accept e]’}”

\(^{10}\) Although focus movement is not a requisite operation in deriving the exhaustiveness reading, it might still be argued that Velleman et al.’s proposal is compatible with the focus approach. For example, we can imagine the morpheme shi to be the overt manifestation of the CLEFT operator in Chinese. Indeed, this is the solution briefly suggested in Erlewine [5], which hypothesizes that shi resides at the clausal spine and takes a propositional argument. Note that, by assuming the role of CLEFT, shi denotes a focus-sensitive operator that values a true proposition against focus alternatives. A focus operator use, thus defined, differs from the copula use in den Dikken’s framework, and needs to be treated as instantiating a separate lexical entry. As such, Erlewine’s (2016) treatment, even when doing away with the movement of cleft phrase, still faces the challenge of addressing the diachronic pattern of coordinated emergence and decline.
As Lin [21] and Li [18], among others, point out, the circumvention of island effects in Chinese might be due to the availability of empty pronouns within complex NPs, suggesting that (11) does not involve a movement-created operator-variable binding relation. This finding is compatible with a biclausal analysis that posits no focus movement, and confirms the diachronic data.

4 Conclusion

This paper lends support to a copular theory of Chinese clefts. In this characterization, clefts are a biclausal structure with a matrix clause headed by a copula verb, and the exhaustive focus interpretation is derived from a presupposition, projected from a definiteness operator within the clausal argument of the copula. Evidence is drawn from a robust diachronic pattern, featuring three converging trajectories in which both the emergence of copular and cleft uses and their decline are coordinated. This pattern, I argue, presents a convincing case for the homogeneity of the copular and the cleft structure. It is thus concluded that historical evidence can be invoked to bear on an ongoing theoretical debate in the field of Chinese syntax and semantics.

Bibliography

The Pragmatics or Speaker Repeat Questions

Gisela Disselkamp

Universität Konstanz, Germany

Abstract. In German, the speaker can repeat a question with an ob-question when the addressee did not reply to the question when it was originally posed. The Table model of Farkas & Bruce (2010) cannot capture this since it assumes ideal discourse with full compliance of both participants. To represent speaker repeat questions in this model, I amend it by overtly taking the speaking and understanding part of the discourse into account.

1 Background

1.1 Speaker Repeat Questions

German direct polar questions have V1-word order, and are embedded with a complementizer ob ‘whether’ and verb-final word order (see Ex. (1)).

(1) a. Regnet es?
   rains it
   ‘Is it raining?’
   b. Paul wei, ob es regnet.
      Paul knows whether it rains
      ‘Paul knows whether it is raining.’

However, in certain contexts the embedded word order can surface without an overt matrix clause under which it is embedded. One of these contexts is the case of speaker repeat questions (see Ex. (2)), as observed by Altmann (1987).

   rains it (no reply) whether it rains
   A: ‘Is it raining?’ B: (no reply) A: ‘Is it raining?/I asked (you) whether it is raining.’

There are a number of restrictions on this context. The speaker of the repeat question may differ from discourse participant who uttered the original question, so long as it is addressed to the same participant as the original question. However, the speaker repeating the question does so not necessarily because he or she is interested in resolving the issue. Other reasons for repeating may be politeness or helpfulness.

Apart from direct questions, the speaker repeat question can also be used to repeat embedded questions, if they were embedded either under a rogative predicate in a declarative sentence, or under a responsive predicate in an interrogative sentence.
Further, a speaker repeat question cannot be used when the question was in fact addressed, i.e. if it was resolved (by an answer) or at least acknowledged (by actual linguistic material, for example Let me think about it or hm ... indicating the addressee is thinking about it). It can be repeated if the addressee asked for a repeat (with a full question like what did you say? or just what?) or otherwise signaled a request for a repeat (for example hm? or just eye contact with a raised eyebrow). If there is no reaction from the addressee at all, the speaker repeat question is also licensed.

The reasons for not answering or asking for a repeat range from not fully understanding or wanting clarification of the content (in this case the addressee will most likely ask for a repeat) up to complete inattentiveness, in which case the addressee didn’t hear the question and may in fact not even be aware that he or she is in a discourse situation.

All of the context restrictions described above hold for polar and constituent questions, as well as embedded questions under declarative rogatives and interrogative responsives. However, since the Table Model which I intend to use so far only represents assertions and polar questions, I will focus on the repeat of a polar question.

1.2 The Table Model

The Table model by Farkas & Bruce (2010) represents discourse by tracking the utterances of the discourse participants A and B. Additionally, the discourse participants public discourse commitments $DC_A$ and $DC_B$ are recorded, which are the sets of propositions that a discourse participant has publicly committed to and which are not yet in the common ground. The common ground $cg$ (the set of propositions accepted by all participants) is also tracked, and possible future common grounds are collected in the projected set $ps$.

Farkas & Bruce (2010) only look at default assertions, default polar questions, and positive and negative responses to the two. If participant A makes an assertion with the meaning $p$, it is added to the Table, to A’s discourse commitments $DC_A$ and the union of $\{p\}$ and the common ground is added to the projected set as a possible future common ground. If A instead utters a question with the denotation $\{p, \neg p\}$, it is only added to the Table and the unions with the common ground are added to the projected set as possible future common grounds. The responses on the other hand always commit the participant to a proposition, i.e. for either response, the denotation of it will be added to the participant’s discourse commitments.

If a proposition is present in every participant’s discourse commitments, it will be added to the common ground and the Table will be cleared of all elements which are entailed by the updated common ground. It is also possible to represent disagreement in this model; if one participant is committed to $p$ and another is committed to $\neg p$, they can ‘agree to disagree’, in which case the Table is cleared

1 Of course, there can be more than two discourse participants.
of all the elements entailed by \( p \) or \( \neg p \), but the discourse commitments are not changed and the common ground will not be updated.

<table>
<thead>
<tr>
<th></th>
<th>Table</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( DC_A )</td>
<td></td>
<td>( DC_B )</td>
</tr>
<tr>
<td>( cg )</td>
<td>( ps = { cg } )</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. The Table Model by Farkas & Bruce (2010)

The aim of the discourse in this framework is to increase the common ground and clear the Table. This is achieved by using the discourse moves described above, where adding to the Table raises issues, and updating the common ground clears the Table.

2 Extension of the Table Model

2.1 Arguments against the Ellipsis Hypothesis

Repeat questions have been analysed as instances of ellipsis of a matrix clause of the form I have asked (cf. Altmann 1987, Zimmermann 2013, Truckenbrodt 2013). While technically, ellipsis would require an over antecedent, i.e. this would only be applicable in cases where the addressee poses a full question for a repeat, recent work by Weir (2013) has shown that ellipsis can also be licensed by an implicit QUD of the form What did X say?. This way the ellipsis analysis can be extended to cover all the cases where the addressee has given some indication that a repeat is wanted. However, this only holds for cases where the addressee was aware of who made the original utterance. If the addressee was unaware of the discourse situation, such a QUD cannot be assumed.

Further, this analysis predicts that in the repeat of an embedded question, the antecedent for the ellipsis process is not the original matrix clause, but the implicit QUD. From the point of view of the ellipsis process, the matrix clause of the previous utterance should be available as an antecedent. On the other hand, from the point of view of the addressee, in cases where he or she didn’t hear the original utterance, the ellipsis site cannot be reconstructed if the antecedent is unavailable. This means that in these cases any analysis of speaker repeat questions as ellipses cannot be motivated.

A second point is that if speaker repeat questions were ellipses, the ellipsis site itself should be available as an antecedent for further anaphoric processes. Both the elided and the non-elided form should allow the same continuations in the discourse. However, only the unelided form can be followed by ach, wirklich? ‘oh, really?’:
The aim of the discourse in this framework is to increase the common ground and clear the Table. This is achieved by using the discourse moves described above, where adding to the Table raises issues, and updating the common ground clears the Table.

### 2. Extension of the Table Model

#### 2.1 Arguments against the Ellipsis Hypothesis

Repeat questions have been analysed as instances of ellipsis of a matrix clause of the form

\[ \text{I have asked} \]

(cf. Altmann 1987, Zimmermann 2013, Truckenbrodt 2013). While technically, ellipsis would require an over antecedent, i.e. this would only be applicable in cases where the addressee poses a full question for a repeat, recent work by Weir (2013) has shown that ellipsis can also be licensed by an implicit QUD of the form

\[ \text{What did X say?} \]

This way the ellipsis analysis can be extended to cover all the cases where the addressee has given some indication that a repeat is wanted. However, this only holds for cases where the addressee was aware of who made the original utterance. If the addressee was unaware of the discourse situation, such a QUD cannot be assumed.

Further, this analysis predicts that in the repeat of an embedded question, the antecedent for the ellipsis process is not the original matrix clause, but the implicit QUD. From the point of view of the ellipsis process, the matrix clause of the previous utterance should be available as an antecedent. On the other hand, from the point of view of the addressee, in cases where he or she didn’t hear the original utterance, the ellipsis site cannot be reconstructed if the antecedent is unavailable. This means that in these cases any analysis of speaker repeat questions as ellipses cannot be motivated.

A second point is that if speaker repeat questions were ellipses, the ellipsis site itself should be available as an antecedent for further anaphoric processes. Both the elided and the non-elided form should allow the same continuations in the discourse. However, only the unelided form can be followed by ‘oh, really?:'

\[ \text{(3) Anna: Ich habe gefragt, ob es regnet.} \]

\[ \text{I have asked whether it rains.} \]

Anna: ‘I asked whether it is raining.’

Ben: Ach, wirklich?

\[ \text{oh really} \]

Ben: Oh, really?’

\[ \text{(4) Anna: Ob es regnet.} \]

\[ \text{whether it rains} \]

Anna: ‘I asked whether it is raining.’

Ben: # Ach, wirklich?

\[ \text{oh really} \]

Ben: Oh, really?’

This observation is independent of whether the addressee asked for a repeat or not.

From this I conclude that speaker repeat questions cannot be analysed as cases of an elided matrix clause. In the following I will focus on cases where the addressee was completely inattentive and did not hear, understand, or acknowledge the previous question.

#### 2.2 The Extension

Ideal discourse conditions include that the speaker and the addressee hear every utterance and are fully attentive. In order to model non-ideal discourse conditions, this must be spelled out overtly. That is, when the speaker makes an utterance, this does not necessarily mean that the addressee hears or fully understands what the speaker said. To capture this, I split up discourse moves into finer parts, using the discourse move-parts *Utter* and *Acknowledge*, as well as the notion of a proposed Table pT\(_X\).

Paralleling the relation of discourse commitments and common ground, this is a separate Table for each participant, which only contains items that are not on the Table yet: utterances made by the participant, or utterances that the participant has already acknowledged as having understood. While *Utter* is the move by which the speaker proposes to add something to the Table, i.e. make it public for discussion, *Acknowledge* is a move by which the other participants signal that they heard the utterance and understood its content. Such signals can be non-linguistic, like eye-contact or nodding, or utterances like *hm.*\(^2\) A reply to an utterance indicates acknowledgement if its content is linked to the content.

\(^2\) Farkas & Bruce (2010) assume that silence after an assertion is agreement. I would amend this to be silence with at least eye-contact.
of the previous utterance in some way. Also, when the speaker is sure of the attention of the addressee(s), he or she may not wait for an acknowledgement but just assume that all utterances are acknowledged.

With this extension, all discourse moves have a threefold structure: an Utter-move that fulfills the licensing/input conditions for the discourse move, a number of Acknowledge-moves, and the completion of the discourse move with context changes as described in Farkas & Bruce (2010), which requires that all discourse participants have acknowledged the utterance. Crucially, Utter always depends on the discourse move it is a part of since it encodes the licensing conditions of that move. Acknowledge on the other hand is the same for all kinds of utterances as it only signals that the content was understood.

When the speaker makes an utterance, he or she proposes to add the utterance to the Table. To achieve this, the utterance is first added to the speaker’s proposed Table. When the other participants acknowledge the utterance, it is also added to their proposed Tables. The condition for the completion of the discourse move is then that the top entries of all participants’ proposed Tables are the same. In this case the discourse move can be completed.

If the acknowledgement is missing, the discourse move cannot be completed. The speaker will take this as a cue that some part or the whole of the utterance was not understood. Especially with respect to questions, which raise issues in this model, the acknowledgement is crucial.

For speaker repeat questions, the Utter-move to be made by participant $A$ toward participant $B$ is then:

a. **Input context conditions:**
   i. $\text{top}(pT_{A,i}) = \langle S[I], \{p, \neg p\} \rangle$
   ii. $\text{top}(pT_{B}) \neq \langle S[I], \{p, \neg p\} \rangle$

b. **Change:**
   $\text{Utter-SRQ}(K_i) = K_o$ such that
   i. $pT_{A,o} = \text{push}(\langle S\{SRQ\}, \{p, \neg p\}, pT_{A,i} \rangle)$

Here, the licensing conditions are that the last utterance was made by $A$ and that it was a question, and that $B$ did not acknowledge this utterance. The actual utterance made is the speaker repeat question, which is added to the speaker’s (here $A$’s) proposed Table $pT_{A}$. The meaning of the repeat question is that of a polar question; namely the set of possible answers.\(^3\)

This move can be visualised in the Table model as performing the change that can be seen between the following two tables:

\(^3\) Probably, there is also not-at-issue meaning here, because repeat questions will likely be different from direct questions in some way, but I am not sure what it is that is not-at-issue.
that move.

Acknowledge on the discourse move it is a part of since it encodes the licensing conditions of

...all utterances are acknowledged. But just assume that all utterances are acknowledged.

...attention of the addressee(s), he or she may not wait for an acknowledgement of the previous utterance in some way. Also, when the speaker is sure of the participants have acknowledged the utterance. Crucially,

Utter changes as described in Farkas & Bruce (2010), which requires that all discourse

...ments, and the completion of the discourse move with context

...-moves, and the completion of the discourse move with context

...can be seen between the following two tables:

<table>
<thead>
<tr>
<th>[S[I], {p, \neg p}]</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>cg</td>
<td>ps = {cg}</td>
</tr>
</tbody>
</table>

**Table 2.** The input context for *Utter-SRQ:* Participant A has uttered a question which was not acknowledged by the addressee B. This is represented by the question being on A’s proposed Table, but not on B’s proposed Table.

<table>
<thead>
<tr>
<th>[S[I], {p, \neg p}]</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S</td>
<td>SRQ], {p, \neg p})</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>cg</td>
<td>ps = {cg}</td>
</tr>
</tbody>
</table>

**Table 3.** *Utter-SRQ* changes the context by adding the new utterance to the speaker’s proposed Table.

This move is (hopefully) followed by an acknowledgement by B of having heard the utterance this time. The general *Acknowledge*-move is

a. **Input context conditions:**

   i. $\text{top}(pT_A) = \langle S[x], ? \rangle$

b. **Change:**

   **Acknowledge**($K_i$) = $K_o$ such that

   i. $pT_{B,o} = \text{push}((S[x], ?), pT_{B,i})$

This means that the utterance must be at the speaker’s proposed Table, and acknowledging simply adds it to the other participants’ proposed Tables. With this move completed, the Table Model will now change to this:

<table>
<thead>
<tr>
<th>[S[I], {p, \neg p}]</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S</td>
<td>SRQ], {p, \neg p})</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>cg</td>
<td>ps = {cg}</td>
</tr>
</tbody>
</table>

**Table 4.** The Table Model after the addressee B has acknowledged the utterance by A by also adding it to B’s proposed Table.
For the discourse move to be completed, the utterance must be added to the Table and change the projected set according to the rules set out in Farkas & Bruce (2010). This is achieved by the following move:

a. **Input context conditions:**
   i. \( \text{top}(p_{TA,i}) = \langle S[1], \{p, \neg p\} \rangle \)
   ii. \( \text{top}(p_{TB}) \neq \langle S[1], \{p, \neg p\} \rangle \)

b. **Change:**
   \( \text{SRQ}(K_i) = K_\alpha \) such that
   i. \( p_{TA,o} = \text{push}(\langle S[\text{SRQ}], \{p, \neg p\} \rangle, p_{TA,i}) \)

This leads to the following Table after the full completion of the discourse move:

<table>
<thead>
<tr>
<th>( \langle S[1], {p, \neg p} \rangle )</th>
<th>( \langle S[\text{SRQ}], {p, \neg p} \rangle )</th>
<th>( \langle S[\text{SRQ}], {p, \neg p} \rangle )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( cg )</td>
<td>( ps = {cg \cup {p}, cg \cup {\neg p}} )</td>
<td>( cg )</td>
</tr>
</tbody>
</table>

**Table 5.** After completion of the discourse move, the utterance is now on the Table and the possible answers are in the projected set as possible future common grounds.

From this point, the discourse can continue in the regular way.

At this point, the speaker repeat question is only licensed in contexts where the addressee does not react in any way to acknowledge the previous utterance. However, if you take asking for a repeat to also signal non-acknowledgement of the content of the previous utterance, these cases will also be covered by the analysis previously presented. This is possible because in the Table model, not all utterances must be moves to add material to the table.

### 3 Conclusion

By extending the Table Model to represent sub-parts of discourse moves, non-ideal discourse can now also be captured in a regular way. The licensing conditions of discourse moves like repeats which are aimed in some way at repairing non-ideal discourse can be represented.

Unfortunately, this model does not allow to track misunderstandings in which both sides are sure they understood correctly and continue to have a conversation. In this case it would be necessary to split the model into separate models for each participant. This is quite impractical as it would mean to make the models inaccessible to the other side and the parts which are public here to be private instead. Instead, I suggest to keep the model as it is and add acknowledged
utterances to the Table in the form that the speaker intended. The addressee who misheard is likely to pick up misheard parts in subsequent utterances, and qua the rules of the model, there would be no common ground updates. The conversation would then progress until one participant notices the inconsistency and makes a clear up the misunderstanding and repair the discourse.

References

An additive ambiguity: summing events or degrees

Cara Feldscher⋆

Michigan State University

Abstract. This paper presents data illustrating a natural class of expressions that is characterized by a systematic ambiguity. Expressions like more, another, additional, and so on can be interpreted as adding degrees or summing eventualities. Rather than positing multiple homophonous morphemes, I provide an analysis that builds the event summing reading from the degree addition reading to account for this ambiguity, using the comparative/incremental more as a case study.

1 Introduction

An understudied use of the English more sums eventualities instead of setting up a comparative. Called “incremental” or “additive” more, this use sums subevents into superevents, adding their measures to make the measure of the superevent (Greenberg, 2010, 2012; Thomas, 2010, 2017). To illustrate, (1) exemplifies the usual comparative use of more, and (2) exemplifies incremental more.

(1) a. I graded three more papers than John.
   b. I am more intelligent than John.

(2) a. Andrew and I have to grade all ten of these papers. He graded six papers, and today I graded three more (papers), so we’re not done. (9 total, 6 for him, 3 for me)
   b. Bill and I are doing a relay race together. He ran two miles, and I ran three more (miles), so we hit the goal of five miles. (5 miles total, 2 for him, 3 for me)

As noted, the truth conditions of (2a) are not that I graded more papers than Andrew did. I only graded three, not three more than he did, as that would contradict the premise here. This means that my grading event of three papers can be added to his paper grading of six papers. For clarity, the truth conditions will be noted for examples, as above.

This use of more has been analyzed as a homophonous morpheme to the comparative. Here, I propose that both readings begin with the comparative

⋆ Thanks to Marcin Morzycki, Alan Munn, Karthik Durvasula, and Yen-Hwei Lin for helpful comments and discussion, as well as to my colleagues at Michigan State University, and to the anonymous reviewers.
morpheme, providing degree addition, and the event summing reading is built from that. In support of this claim, I provide evidence that several other English words pattern similarly, displaying this same ambiguity, which I use to delineate the natural class of additive expressions in Section 2. In 3, I review first the relevant theory of the comparative, and the theory that I will be using here, and then I review what has been said about this data in the literature. From there, in Section 3.2 I provide a denotation for a morpheme that combines with additive expressions to create the event summing readings. Section 4.2 takes that and builds a denotation for an event-summing version. Section 5 discusses limitations and extensions of the analysis. Finally, I conclude in Section 6.

2 Defining a natural class

The standard comparative use of more can be rephrased to be about addition. I am more intelligent than John means I have John’s degree of intelligence plus some. I graded three more papers than John means that I graded the number he graded, plus three. The relevant aspect of the comparative here is degree addition. In contrast, in the event summing reading, more primarily sums events or states, instead of adding degrees. If summing two events of paper grading for example, as in (2a), the measures of the two subevents are necessarily added to get the measure of the super event. So the event summing reading does involve adding degrees, but as a result of summing events. Here I provide data indicating that several other words and phrases in English have this exact same ambiguity.

To start with, Thomas (2011) observed what they called “additive another”, which sums eventualities, given the right QUD conditions, shown in (4). Another can also be used for degree addition, as in (3), similar to comparative more.

(3) Another – degree addition
   a. I can run much further than Bill can. He can run two miles in one go, but I can run another three. (I can run 5 miles at once)
   b. I thought I could only grade two papers in one sitting, but I managed to grade another three. (5 in one event).

(4) Another – event summing
   (A: How wide are those two pieces of furniture together?)
   B: The cabinet is 4 ft wide. The shelves are another 3 ft wide. (3ft shelves, 7ft furniture total)

At this point, it can be seen that another patterns with more in having a degree addition reading where it adds two degrees to get the measure of one eventuality, and an event summing reading where it sums two eventualities to make a super-eventuality.

Other words that follow this pattern include extra and additional, shown in (5-8). Other additive expressions are more complex phrases, such as on top of or in addition to, exemplified in (9-10), but these show the same ambiguity.
(5) “Extra” – degree addition
   a. I only ordered ten dumplings, but the box they delivered has an extra two. (12 in box)
   b. I only planned to run one mile, but then I ran an extra two today. (3 in one event)

(6) “Extra” – eventuality summing
   a. We have a dozen dumplings, since 10 are on the plate, and the box has an extra two. (2 in box, 12 total)
   b. I ran one mile yesterday, and then ran an extra two today, so I hit my goal of three miles this week. (2 today, 3 total)

(7) “Additional” – degree addition
   a. I only ordered ten dumplings, but the box they delivered has an additional two. (12 in box)
   b. I only planned to run one mile, but then I ran an additional two today. (3 in one event)

(8) “Additional” – eventuality summing
   a. We have a dozen dumplings, since 10 are on the plate, and the box has an additional two. (2 in box, 12 total)
   b. I ran one mile yesterday, and then ran an additional two today, so I hit my goal of three miles this week. (2 today, 3 total)

(9) “On top of” – degree addition
   a. I only ordered ten dumplings, but the box they delivered has two (more) on top of that. (12 in box)
   b. I only planned to run one mile, but then I ran two (more) on top of that today. (3 in one event)

(10) “On top of” – eventuality summing
    a. We have a dozen dumplings, since 10 are on the plate, and the box has two (more) on top of that. (2 in box, 12 total)
    b. I ran one mile yesterday, and then ran two (more) on top of that today, so I hit my goal of three miles this week. (2 today, 3 total)

To conclude this section, I have provided data showing that several words or phrases in English show the same ambiguity that more does, supporting analyzing these as a natural class of additive expressions. In the following section, I describe the background in the literature for both comparative and event-summing more.

3 Background

3.1 Comparative more

Much more has been written about the comparative than will be covered here. I am confident that my analysis could be adapted to work in more than one
theory of the comparative, and the data given in this paper does not suggest any reasons for choosing one particular structure over the other. Given this, I will only review only the analysis I will be using, that of Wellwood (2015).

Wellwood (2015) breaks down comparative more into much-er (Bresnan, 1973), where much provides a contextually supplied measure function \( \mu \) via an assignment function \( A \). She argues that a measure function must be accessible that measures individuals and eventualities, using a scale that is monotonic on the eventuality or individual. Schwarzschild (2006) shows that measure phrases must be monotonic on the mereology of the event or individual being measured in the case of partitives, or else it will be ungrammatical, as in (11).

\[
\text{(11) * Bill poured 50 degrees of coffee.}
\]

Wellwood suggests that even so, there are still multiple scales possible. In (12) and (13), more must be able to access multiple scales in the same construction.

\[
\text{(12) John poured more coffee than Bill did, in terms of } \{ \text{volume, weight, number of cups filled} \} \\
\text{(13) John ran more than Bill did, in terms of } \{ \text{distance, time, number of races ran} \}
\]

Here are the denotations Wellwood proposes in (14), where applying much to -er produces the familiar more, and where -er is flexible in the type of measure function that it takes in.

\[
\text{(14)} \\
\begin{align*}
\text{[-er]} &= \lambda g(\alpha, d) \lambda d \lambda \alpha. g(\alpha) > d \\
[\text{much}_\mu]^A &= A(\mu) \\
[\text{more}_\mu]^A &= \text{[-er]}(\text{[much]}) \\
[\text{more}_\mu]^A &= \lambda d \lambda \alpha. A(\mu)(\alpha) > d
\end{align*}
\]

I build my analysis using the denotation of more (much-er) from this paper for two reasons. First, it is a reasonably straightforward denotation, and it is simpler to build a theory without movement to start with. Second, as has already been demonstrated in the data provided so far, the ambiguity occurs in constructions measuring different types of things. It makes sense to work with a theory designed to capture that fact. However, I will use this simplified version in (15) and omit the assignment function, to keep things clearer with fewer variables and parentheses.

\[
\text{(15)} \\
[\text{more}_\mu] = \lambda d \lambda \alpha. \mu(\alpha) > d
\]
3.2 Incremental more

The main requirements for a denotation of event summing *more* are (1) the local degree must measure the local event (assertion), (2) there must be a non-local event that can be measured and can be summed with the local event, and (3) the measure of the super event is the sum of the measure of the local and non-local events. There are two accounts in the literature of what they call “incremental” *more* or “additive” *more*, respectively, by Thomas (2010) and Greenberg (2010, 2012), where they analyze the event summing *more* as a separate morpheme from comparative *more*. I will refer to both here as incremental *more*, to avoid any confusion with “additivity” possibly referring to either adding degrees or summing events.

Both authors analyze incremental *more* as a homophonous morpheme to comparative *more* because they have different translations multiple languages. Incremental *more* is often translated as the same morpheme as *still*, but comparative *more* is translated using a different word. The example here is in Hebrew, from Greenberg (2012).

(16) Rina od yeSena
Rina still asleep
‘Rina is still asleep.’

(17) (etmol axalti 3 tapuzim) ha-yom axalti od (tapuzim)
yesterday I-ate 3 oranges) the-day I-ate od (oranges)
“(Yesterday I ate 3 oranges.) Today I ate some more (oranges).”

(18) hayom dani r’ayen SloSa studentim. etmol hu r’ayen
today Danny interviewed three students. yesterday he interviewed
yoter/#od yoter/od
‘Today Danny interviewed three students. Yesterday he interviewed more (than three).’

Greenberg (2010, 2012) contributes a denotation that thoroughly ensures that there is a previous event and it can be summed with this eventuality to create a super-event. In order for events to be summable, they must both be stages of the same super-event. That is, the summed event must be a more developed version (Landman, 1992). Thomas (2010) presents a denotation that accomplishes much the same thing, but with the variation that the non-local event must be in the alternative set of relevant eventualities. This captures how different types of events can be summed, as shown in (19).

(19) A: How much did you exercise last week?
B: I ran for two hours and I biked for three hours more.

---

1 I will refer to the “local” event or degree to mean the one in the sentence, near to the additive expression. “Non-local” then refers to the other event mentioned, but non-locally meaning further from the additive expression and possibly in another sentence. The “summed event” is these two events added.
For his denotation of additive *another* in Thomas (2011), Thomas introduces the QUD as a use condition, as previously mentioned. While yet another way to model the presupposition of the alternative event, this still requires assuming another homophonous morpheme.

(20) a. A: How wide are those two pieces of furniture together?
   B: The cabinet is 4 ft wide. The shelves are another 3 ft wide. (3ft shelves, 7ft total)

b. A: How wide are the shelves?
   B: The shelves are (*another) 3 ft wide.

Both Greenberg and Thomas have analyses that account for the event summing reading of *more* by positing a homophonous morpheme, and Thomas similarly accounts for event-summing *another*. What they do not account for, however, is the behavior of additive expressions in English as a whole, where expressions with degree addition readings also have event-summing readings.

Thomas (2017) more recently takes on a slightly different homophony, looking at comparison, event summing, and continuation (as expressed by *still*). They observe that cross-linguistically, there are languages where comparison is homophonous with event summing (like English), languages where event summing is homophonous with continuation (like Hebrew), and languages where all three are homophonous (Romanian for example), but no languages in their study where comparison is homophonous with continuation and not event summing. They develop a theory in Distributive Morphology to account for the homophony. What all three of these meanings have in common in their theory is a rising scale segment (Schwarzschild, 2012), so in languages like Romanian, homophony is achieved by assuming a Lexical Insertion rule that applies everywhere there is the Deg head RISE. This is a rising scale segment, and in a comparative, it describes the difference between the two things being compared. Similar to the theory I will propose here, a functional head ADD applies higher up and sets it such that the rising scale segment describes the increase in the measure of the summed event, making it event summing. Thomas goes on to create continuation (*it is still raining*) by adding another functional head CON to this projection. The patterns of homophony in different languages is then achieved by assuming different Lexical Insertion rules in different languages, where the three meanings are spelled out with one, two, or three different words.

While this fits the cross-linguistic pattern very nicely, in this paper I focus on a related, but slightly different issue. Specifically, here I wish to address the issue of a whole class of additive expressions featuring this ambiguity. To that end, I will pursue a slightly different analysis also building event summing from degree addition, focusing on *more*, but with the intention that this analysis can be extended to the rest of the class of additive expressions.

### 3.3 Additive *again*

One other place where “additivity” has been discussed is with additive *again*, a use of *again* that deals with degrees instead of the usual repetition or restoration
of eventualities (Feldscher, 2016). The availability of this use of *again* varies by dialect, and it additionally has a very restricted distribution, so it seems unlikely that an analysis of additive *again* will wholly parallel that of additive *more*, but some points will be reviewed here. An example of the usual use of *again*, showing repeating an event or restoring a state, is in (21), and “additive”, or degree-adding *again*, is in (22).

(21) I shut the door again.

(22) These apples are half again as expensive as those oranges.

\[ \text{apples} 1.5 \times \text{oranges} \]

There is some variation on what the structure of a sentence with *again* can look like. The British word order is the (b) version in (23). Some examples with other factor phrases or versions without factor phrases are shown in (24).

(23) a. Neville is half *again* as tall as Pansy. \(1.5 \times \text{Pansy's height}\)
    b. Neville is half as tall *again* as Pansy. \(1.5 \times \text{Pansy's height}\)

(24) a. Neville is a *third* again as tall as Pansy. \(1.33 \times \text{Pansy's height}\)
    b. Neville is *three quarters* again as tall as Pansy. \(1.75 \times \text{Pansy's height}\)
    c. Pansy is four feet tall. Neville is *that* again. \(2 \times \text{Pansy's height}\)
    d. Brazil nut trees...are emergent species, half as tall again as most canopy trees. (Example from the OED, meaning 2x)

The “default” use of *again* is the one that repeats or restores events or states. However the degree-addition use of *again* has no presupposition of a prior eventuality. The sentence in (22) is felicitous if the speaker has never priced apples before, requiring only that there be another degree to compare it with. Feldscher (2016) defines a denotation for *again* but it had to be a different morpheme from the regular repetitive *again*, in order to incorporate degrees.

Feldscher (2016) hypothesizes that the reason *again* is dialectal is that it is backformation, rather than a normal part of the class of additive expressions with their ambiguity. The dialectal nature of this would make sense if standard *again* deals with events, additive expressions deal with degree addition, and additive expressions achieve an event summing reading by combining with some other morpheme. So based on this data, I conclude that the ambiguity should be built compositionally by assuming additive expressions to denote degree addition, and assuming there to be a morpheme that, when combined with a degree-addition morpheme, results in event summing. This is also suggested in that event summing also adds degrees, but not vice versa.

4 Analyzing the ambiguity of *more*

In this section, I describe *more* as a case study for the rest of the additive expressions, as the comparative is a well-studied phenomenon, giving this analysis
an easier starting point. Given the robust ambiguity in the natural class of additive expressions, however, I suggest that both readings should build from the same morpheme, as opposed to a homophonous event summing morpheme as suggested by Greenberg and Thomas. As the event summing reading does involve degree addition in a way, I suggest an analysis where the *more* morpheme basics are that of degree addition, and then a separate morpheme applies to that one in order to derive the event summing reading.

4.1 Viewing comparative *more* as degree addition

As previously mentioned, the comparative can be thought of as involving addition. Below is the denotation adapted from Wellwood’s denotation for the basic comparative repeated in (25), and then the change I am suggesting in (26), where it is rephrased to be about addition. The comparative is usually thought of as saying $A > B$. The change I propose is mathematically equivalent: $A > B$ iff $A = B + C$, where $C$ is some positive, nonzero number.

\begin{equation}
[\text{more}_p] = \lambda d \lambda \alpha. \mu(\alpha) > d
\end{equation}

\begin{equation}
[\text{more}_p] = \lambda d \lambda \alpha \lambda d'. \mu(\alpha) \geq d + d'
\end{equation}

Another argument in favor of reconsidering the typical comparative morpheme to involve addition comes from differential comparatives, which explicit reference a differential degree (Von Stechow, 1984). For example, *Neville is three inches taller than Pansy*. This means that $d_N \geq (d_P + 3)$. The presence of two local degrees means that the differential version of the comparative requires addition. This minimally changes the denotation for a differential *more* to the denotation in (27), which calls for two degrees locally and adds them.

\begin{equation}
[\text{more}_p] = \lambda d \lambda d'. \lambda \alpha. \mu(\alpha) \geq d + d'
\end{equation}

Schwarzchild and Wilkinson (2002) make an argument to unify the comparative morpheme used in differentials (requiring two degrees) with the one used in cases with only one local degree. They follow the same mathematical reasoning that $A > B$ iff $A = (B + \text{something})$. Working in an interval-based semantics to express this intuition, they propose a covert *SOME* which expresses the “some positive amount” part of this intuition. In non-differential comparatives, instead of the overt differential degree, *SOME* appears instead. This allows the differential and non-differential comparatives to both call for two degrees / intervals, making the two instances of the comparative *more* unifiable to one morpheme. This gives the two comparatives a parallel expressed in (28) and (29).

\begin{equation}
\text{Neville is 10 inches taller than Pansy.} \\
\quad d_N \geq d_P + 10
\end{equation}

\begin{equation}
\text{Neville is *SOME* taller than Pansy.} \\
\quad d_N \geq d_P + d \text{ where } d \text{ is a positive, nonzero degree.}
\end{equation}
Because it matches the intuition that the comparative involves degree addition, I follow Schwarzchild and Wilkinson in my analysis in unifying both versions of the comparative to one morpheme, and assuming something like their covert SOME applies with non-differential cases. Thus from this point, the denotation for more that I will be working with is the differential one from (27).

4.2 Deriving event summing

In this section I build the event summing reading from the degree addition reading, in order to explain why English at least has a natural class of additive expressions, where if they have a degree addition reading they also have the event summing reading. In order to fully explain this robust ambiguity, I posit that a separate morpheme, what I’m calling sum, builds event summing from a morpheme with degree addition, instead of each additive expression having a degree addition morpheme and an event summing morpheme.

While this example below attempts to push the event summing reading as the intended meaning by stating a five mile goal, the following sentence is still ambiguous, and a degree addition reading is still possible.

(30) Bill and I are doing a relay race together. He ran two miles, and I ran three more, so we hit the goal of five miles. ✓ (5 miles total, 3 for me) ✓ (more than the goal, 7 miles total, 5 for me)

To make an unambiguous sentence, the addition of a standard degree in an overt than-phrase blocks the event summing reading, as in (31b).

(31) Bill ran two miles...
   a. and I ran three mile more. (I ran 5 OR 3)
   b. and I ran three miles more than Bill ran. (I ran 5 NOT 3)

Based on this data, I posit that the than-phrase structurally blocks the sum morpheme, blocking the event summing reading. In (32) I show Wellwood’s tree for comparative more, adapted to the sentence from (31). In (33), the same structure appears, but instead of a than-phrase, there is sum.

(32) Degree addition (Wellwood, 2015)

```
VP
   (v, t)
run
   ⟨d⟩
3-miles
   Deg
   ⟨d⟩
3-miles
   Deg'
   ⟨d⟩
more
   thanP
```

(33) Degree addition (Wellwood, 2015)

```
VP
   (v, t)
run
   ⟨d⟩
3-miles
   Deg
   ⟨d, ⟨d, ⟨v, t⟩⟩⟩
more
   thanP
```

112
As a reminder, the event summing reading requires (1) the local degree must measure the local event (asserted), (2) there must be a non-local event that can be measured and can be summed with the local event, and (3) the measure of the super event is the sum of the measure of the local and non-local events. This must be accomplished by applying the sum morpheme to more. The local event is asserted, as is the summed event. To model the non-local event, I choose, like Thomas (2017), to model it with an index like a pronoun in the discourse. Like the contextually-supplied measure function of Wellwood (2015), sum calls for a contextually supplied eventuality v from the discourse. Based on these pieces, the denotation I propose for the event summing morpheme sum is in (34).

\[
\text{sum}_v = \lambda f \langle d, \langle v, t \rangle \rangle \lambda d \mu v' \mu(v') = d \land f(\mu(v))(d)(v \oplus v')
\]

Context supplying v, sum takes in a thing of type more, which produces the second line in (34). sum builds in the event summing with the \( \oplus \) operator, but degree addition from more contributes the part of the truth conditions that the addition of the measures of the subeventualities make the measure of the supereventuality. The two conjuncts of this morpheme are that the local eventuality \( v' \) measures \( d \), and that the measure of summed eventualities is the measure of the local eventuality \( d \) plus the measure of the non-local eventuality \( v \).

To continue the sample computation, it should be clearer to assume an eventuality for \( v \) in the computation. In (35), information for the contextually relevant non-local event \( v \) is included, but underlined to set it off from the rest of the computation, as it is not actually present. In this example, if we’re talking about how Bill ran the first 2 miles of the 5 mile relay, and I ran three more, we can assume the underlined parts:

\[
\text{sum}_v \text{ more}_v = \lambda d \mu v' \mu(v') = d \land \mu(v_{B\text{-run}} \oplus v') = d + 2\text{-mi}
\]
\[
\text{sum}_v \text{ more}_v = \lambda v' \mu(v') = 3\text{-mi} \land \mu(v_{B\text{-run}} \oplus v') = 3\text{-mi} + 2\text{-mi}
\]
\[
\text{run \: 3-miles \: sum}_v \text{ more}_v = \lambda v' [\text{run}(v') \land \mu(v') = 3\text{-mi} \land \mu(v_{B\text{-run}} \oplus v') = 3\text{-mi} + 2\text{-mi}]
\]

At the end of this computation, the truth conditions are that the local event is one of running, it measures three miles long, and that this event summed with
the contextually relevant non-local event measures 5 miles long. Thus it reaches the correct truth conditions, and the event summing reading is directly enabled by the presence of degree addition in the morpheme that SUM applies to.

5 Extensions

This analysis is tailored for more, and crucially it relies on the SUM morpheme applying where the than-phrase applies to otherwise disambiguate. However, a difference between more and the other additive expressions is that more takes a than-phrase and they mostly do not. One other phrase, on top of, requires two local degrees, and has an of-phrase possibly like the than-phrase, but unlike the than-phrase, both local degrees are necessary and can’t be omitted.

(36) a. I ordered ten dumplings, but they gave me another two *than / ? from / *of what I ordered.
   b. I ordered ten dumplings, but they gave me an extra two *than / ? from / *of what I ordered.
   c. I ordered ten dumplings, but they gave me an additional two *than / ? from / *of what I ordered.
   d. I ordered ten dumplings, but they gave me two on top *(of) what I ordered.

A fully fleshed-out account of the additive ambiguity should account for all the ambiguity, not just the ambiguity in the most well-researched instance of it. Therefore, an important next step is to extend the account to other additive expressions. However, the groundwork is laid for sum to build event summing from morphemes with degree addition.

A different unsolved puzzle is that the addition of an unstressed some, as in (37) unambiguously gets the event-summing reading.

(37) *Yesterday I ran 2 miles, and today I ran some more.

It is unclear what exactly “some” is here. The homophony with sum is almost certainly accidental, as it can not occur with an overt measure phrase.\(^2\)

(38) *Yesterday I ran 2 miles, and today I ran some 3 miles more.  
(\textit{Intended: ran 3 today, *5 today})

One possibility is it is Schwarzschild and Wilkinson’s SOME, not SUM, in which case the question still arises as to why this forces the event-summing reading. Or if this “some” is yet something different, we still can ask the question of why the even-summing reading is blocked here.

Finally, a remaining question is how this relates to cross-linguistic data. As previously discussed, Thomas (2017) aims to capture a different generalization

\(^2\) Ignoring the “approximately” reading of some. The sentence in (38) is grammatical for that reading, but that is a different use of some than in (37).
than the one here, which is the cross-linguistic trend of homophony between comparison (more), event summing (some more), and continuation (still). Here, I attempt to instead capture the generalization of a class of expressions with an ambiguity between degree addition and event summing, starting with the comparative. The syntax is still limited to the comparative, but the groundwork is laid for morphemes with degree addition to create event summing readings when combined with sum. Following Thomas (2017), the next step from there is to see how this pattern fits in with the cross-linguistic three-part trend. Given still and again, there must be some connection to continuation, or possibly to aspect, which this analysis does not currently convey.

6 Conclusion

Here I have presented data showing that there is a natural class of additive expressions in English showing an ambiguity between a reading where they add degrees, and a reading where they sum events. I offer an analysis building the event summing reading from the degree addition capability, using more as a case study, discuss what this shows about the nature of how we conceptualize addition, and conclude by confronting the limitations of this theory.

References

Restructuring and Actuality Entailment in Mandarin Obligatory Control

Yuyin He
Harvard University

Abstract. This paper discusses actuality entailment in Mandarin obligatory control predicates taking a TP complement clause. Actuality entailment occurs when perfective aspect markers are embedded in the complement clause of a restructuring control predicate. I argue that actuality entailment derives from the aspect marker scoping over a single event yielded by size restructuring. Via the Principle of Event Identification across Worlds, we obtain one event in the actual world and in all of the subject’s desire worlds.

1 Introduction

Many languages with a morphological distinction between perfective and imperfective aspect show actuality entailments with root modals. Namely, perfective aspect on a root modal indefeasibly infers that the proposition expressed by the modal’s complement took place in the actual world. The examples from Italian [?] and Greek[?] show that denying the complement clause in the actual world yields a contradiction for sentences in (1a) and (2a) with a perfective aspect on a root modal, while those with an imperfective aspect in (1b) and (2b) remain natural.

(1)  a. Gianni ha potuto parlare a Maria, #ma non lo ha fatto.
    Gianni can-pst-pfv talk to Maria but not it do-pst-pfv
    ‘Gianni was able to talk to Maria, #but he didn’t do it.’

   b. Gianni poteva parlare a Maria, ma non lo ha fatto.
    Gianni can-pst-impf talk to Maria but not it do-pst-pfv
    ‘Gianni was able to talk to Maria, but he didn’t do it.’

   Italian, Hacquard (2008:2)

(2)  a. Boresa na tu miliso #ala òen tu
can.pst-pfv:1s NA him talk.non-pst-pfv:1s but NEG him
     talk.pst-pfv
     ‘I was able to talk to John, #but I did not talk to him.’

   b. Borusa na sikoso afto to trapezi ala òen to
    can.impfv:1s NA lift.non-pst.pfv:1s this the table but NEG it
    lift.impfv
‘(In those days), I could lift this table but I didn’t lift it.’

Greek, Bhatt (1999)

Similar phenomenon also has been observed for restructuring predicate \textit{volere} ‘want’ in Italian\cite{Bhatt1999}, as we can see in (3).

\begin{enumerate}
\item[(3)] a. Gianni ha voluto parlare a Maria, #ma non lo ha fatto.
\hspace{1cm} Gianni want-pst-pfv talk to Maria but not it do-pst-pfv
\hspace{1cm} ‘Gianni wanted to talk to Maria, #but he didn’t do it.’
\item b. Gianni voleva parlare a Maria, ma non lo ha fatto.
\hspace{1cm} Gianni want-impf talk to Maria but not it do-pst-pfv
\hspace{1cm} ‘Gianni wanted to talk to Maria, but he didn’t do it.’
\end{enumerate}

Interestingly, though perfective/imperfective aspect particles never attach to modals in Mandarin, actuality entailment exists in some control constructions where the control verb is a restructuring predicate. Different from the case for Italian \textit{volere} ‘want’, Mandarin obligatory control constructions (MOC henceforth) allow perfective aspect particles attached to either the matrix control predicate or the predicate in the controlled complement.

However, actuality entailment only occurs when the perfective aspect particle is on the complement predicate\cite{Bhatt1999}. The perfective marker \textit{le} and experiential marker \textit{guo} are two perfective markers in Mandarin. \textit{Le} and \textit{guo} in the complement clause force the proposition expressed by that clause to hold in the actual world, indicated by the ungrammaticality of denying the complement clause. However, the same aspectual markers on the matrix predicate \textit{qing} ‘invite’, \textit{bi} ‘force’ do not yield such actualization of the complement. The same pattern also holds for other control predicates such as \textit{quian} ‘persuade’.

\begin{enumerate}
\item[(4)] a. wo qing ta chi-le/guo fan, #keshi ta mei lai.
\hspace{1cm} I invite he eat-PRF/EXP rice but he NEG come
\hspace{1cm} ‘I had invited him to eat, #but he had not come.’
\item b. wo qing-le/guo ta chi-fan, keshi ta mei lai.
\hspace{1cm} I invite-PRF/EXP he eat-rice but he NEG come
\hspace{1cm} ‘I had invited him for dinner, but he had not come.’
\item[(5)] a. Zhangsan bi Lisi chi-le/guo fan, #keshi ta mei chi.
\hspace{1cm} Zhangsan force Lisi eat-LE/EXP rice but he NEG eat
\hspace{1cm} ‘Zhangsan had forced Lisi to eat, #but Lisi didn’t eat.’
\item b. Zhangsan bi-le/guo Lisi chi fan, keshi ta mei chi
\hspace{1cm} Zhangsan force-PRFEXP Lisi eat rice but he NEG eat
\hspace{1cm} ‘Zhangsan had forced Lisi to eat, but Lisi didn’t eat.’
\end{enumerate}

In the following discussion, section 1 will present size restructuring in MOC and argue that in a restructuring construction, Aspect phrase is not projected in the complement clause. Section 2 talks about the semantic assumptions. Section 3 introduces the analysis. Section 4 concludes the paper.
2 Aspect in MOC

2.1 Size Restructuring and MOC

The picture of Mandarin control constructions is unclear due to the lack of case and tense morphology as well as any formal complementizer. Following [?], I propose that MOC involves size restructuring. [?] and [?] argue that cross-linguistically the thematic domain $\Theta$ is first constructed, then an inflectional domain $\Phi$ is added and lastly an operator domain $\Omega$ closes off a clause. Size restructuring arises if not all domains are projected. This yields two types of size restructured complements: infinitives that are just $\Theta$-domains and infinitives that are $\Phi$-domains.

\[
\begin{array}{c}
\Omega \ (\text{CP}) \\
C \quad \Phi \ (\text{TP}) \\
T/\text{Mod} \quad \Theta \ (\text{vP})
\end{array}
\]

Similarly, Mandarin has three types of complement clauses: CP, TP\(^1\) and vP. However, only predicates selecting TP or vP complement requires an obligatory interpretation between the covert embedded subject and the matrix argument as a controller.\(^2\) These are predicates which I termed as MOC predicates involving size restructuring. For instance, predicates such as *shuo* ‘say’, *gaosu* ‘tell’ in (7) are able to project an overt subject, no matter if it corefers with a matrix argument or not. More evidence from [?] shows that they are able to project auxiliaries and independent aspect markers in the embedded clause as well. Hence I assume that these predicates can take a full CP as complement. Predicates in (8) and (9) are unable to project an overt subject in the complement clause, hence they select a clause size that is smaller than those in (7) as complement. However, as pointed out by [?], [?], [?] among others, predicates such as *dasuan* ‘plan’, *bi* ‘force’ in (8) are optionally able to project semantically-compatible auxiliaries. Hence I assume they select a TP complement clause but predicates such as *shefa* ‘try’, *kaishi* ‘begin’ in (9) select a vP complement clause, which is too small to project functional layers containing auxiliaries and aspect markers (c.f. [?])

\[
\begin{align*}
\text{(7)} \quad &a. \ \text{Zhangsan} \quad \text{shuo} \quad \text{[ta}\_j/Lisi \ \text{chi} \ \text{le} \ \text{fan}] \\text{.} \\
&\text{Zhangsan say he/Lisi eat prf food}
\end{align*}
\]

\(^1\) The label TP here corresponds with any syntactic projection which is larger than vP but smaller than CP ($\Phi$ domain in [?] and [?]). It does not necessarily assume that Mandarin has tense phrase syntactically projected. Actually, which syntactic layer is able to project in a TP clause is still not clear. I will leave it for future research.

\(^2\) I doubt that some of the diagnostics for monoclauasality in previous work such as focus-fronting, Negative Polarity Items are constrained by semantic factors rather than syntactic constraints that are sensitive to clause boundary. Hence I focus more on the restrictions on covert subjects, which I believe to be the core of control theory since its birth.
‘Zhangsan, says that he/Lisi has eaten.’

b. Zhangsan, gaosu Lisi [ta_i/k/Wangwu] le food
Zhangsan tell Lisi he/Wangwu eat PRF food

‘Zhangsan, told Lisi, that he/Lisi/Wangwu has eaten.’

(8) a. Zhangsan, dasuan [(ta_i/*Lisi) qu Beijing].
Zhangsan plan he/Lisi go Beijing
Intended: ‘Zhangsan planned to go to Beijing.’/*Zhangsan planned that Lisi will leave Beijing’

b. Zhangsan, bi Lisi [(ta_i/*Wangwu) zaodian lai].
Zhangsan force Lisi he/Wangwu earlier come
Intended: ‘Zhangsan forced Lisi to come earlier.’

(9) a. Zhangsan, shefa [(ta_i/*Lisi) changge].
Zhangsan try *he/*Lisi sing
‘Zhangsan tried to sing’

b. Zhangsan, kaishi [(ta_i/*Lisi) changge].
Zhangsan begin *he/*Lisi sing
‘Zhangsan began to sing.’

The complement clause size has been argued to have a direct correspondence with partial control and exhaustive control by [?]. However, the predicates in (??) are in a syntactically monoclausal construction [?], but they are partial control predicates because both sentences below are compatible with collective adverb yiqi ‘together’. Hence the distinction between partial control and exhaustive control does not necessarily relate to the distinction between biclausality and monoclausality.

(10) a. Zhangsan, bi Lisi PRO_i+j yiqi chi fan.
Zhangsan force Lisi together eat rice
‘Zhangsan forces Lisi to have a meal together.’

b. Zhangsan, qing Lisi, PRO_i+j yiqi chi fan.
Zhangsan invite Lisi together eat rice
‘Zhangsan invite Lisi to have a meal together.’

Given that the MOC predicates taking a vP complement involve various types of predicates with very different denotations [?], I will focus on control predicates taking a TP complement in the following discussion.

### 2.2 Aspect in the Complement Clause of MOC

In a restructuring construction, there is no aspect projection in the complement clause of MOC. Based on Neg-raising phenomena in Mandarin, [?] has shown the following structure of aspect and negation projection in (??). The negative particle mei can only project above Aspect phrase.
The data below in (??) from [?] has shown that the complement clause of *bi ‘force’* does not contain its own projection of Asp that could host the embedded negation of *mei* in (12d). According to [?], for a sentence in (12a), it is possible to host negative particle *bu* in the embedded clause like (12b) does. However, when negation and perfective aspect co-occur, the perfective allomorph *you* must appear overtly at the matrix level and the negative particle surfaces as *mei* rather than *bu*. Namely, even the aspect marker attaches to the embedded predicate, the embedded clause does not contain an aspect projection. [?] hence suggests that the sentences in (12) are monoclausal in the sense that *bi ‘force’* is an inflectional-layer functional head. [?] also show a similar analysis to Italian *volere*.

(12)  
a. Zhangsan bi Lisi chi-le fan.  
Zhangsan force Lisi eat-PRF food  
‘Zhangsan made Lisi have a meal.’
b. Zhangsan bi Lisi bu chi fan.  
Zhangsan force Lisi NEG eat food  
‘Zhangsan forced Lisi not to have a meal.’
c. Zhangsan mei-you bi Lisi chi fan.  
Zhangsan NEG-PRF force Lisi eat food  
‘Zhangsan had not forced Lisi to have a meal.’
d. * Zhangsan bi Lisi mei-you chi fan.  
Zhangsan force Lisi NEG-PRF eat food  

Furthermore, sentences in (13) has shown that before aspect-marker lowering to the embedded verb, the embedded clause is still able to host VP negation *bu*, while once the aspect-marker ‘lowers’ to the embedded clause, the embedded verb phrase is unable to host embedded negation any more.
verb phrase is unable to host embedded negation any more. While once the aspect-marker ‘lowers’ to the embedded clause, the embedded verb, the embedded clause is still able to host VP negation.

suggests that the sentences in (12) are monoclausal in the sense that bi-predicate, the embedded clause does not contain an aspect projection. Namely, even the aspect marker attaches to the embedded clause like (12b) does. According to [13] also show a similar analysis to Italian

However, when negation and perfective aspect co-occur, the perfective allomorph is possible to host negative particle

In (12d). According to [13] notices that temporal reference of Mandarin complement clause is largely constrained by the lexical semantics of the matrix verb and the constraint cannot be overruled by the use of an aspect marker. He shows two groups of predicates. The group of predicates impose a fixed temporal reference upon their complement clause even though the complement clause does not contain temporal adverbials or aspectual markers. Among them are qiangpo ‘force’, jianyi ‘suggest’ in (14), namely the restructuring control predicates taking a TP complement in this paper. On the contrary, another group of predicates such as shuo ‘say’ and renwei ‘believe, think’, namely the CP complement predicates, do not impose such fixed temporal relation, as illustrated in (15).

How to analyze the status of the control predicate here is a controversial topic, which I will leave for future research. However, the interaction between aspect and negation shows one important thing: Due to restructuring, even the control construction is monoclusal, the complement clause size in (13b) gets truncated more (without any possibility of VP negation) and the control predicate combines with a predicate of event, forming another complex event.

Along with the truncated structure, restructuring predicates also show a difference in temporal restrictions upon their complements from those of non-restructuring predicates. [?2] notices that temporal reference of Mandarin complement clause is largely constrained by the lexical semantics of the matrix verb and the constraint cannot be overruled by the use of an aspect marker. He shows two groups of predicates. The group of predicates impose a fixed temporal reference upon their complement clause even though the complement clause does not contain temporal adverbials or aspectual markers. Among them are qiangpo ‘force’, jianyi ‘suggest’ in (14), namely the restructuring control predicates taking a TP complement in this paper. On the contrary, another group of predicates such as shuo ‘say’ and renwei ‘believe, think’, namely the CP complement predicates, do not impose such fixed temporal relation, as illustrated in (15).

(13) a. Zhangsan mei bī guò Lisi bu chǐ-fàn.
   Zhangsan NEG force EXP Lisi NEG eat-food
   ‘Zhangsan had not forced Lisi not to eat.’

b. * Zhangsan mei bī Lisi bu chǐ guò fán.
   Zhangsan NEG force Lisi NEG eat EXP food
   ‘Intended: Zhangsan had not forced Lisi not to eat.’

(14) e₁ < e₂ (matrix event precedes complement event)

Ta qiangpo/jianyi wo kào dàxué.
he force/suggest I take-exam university

‘He forced me to/suggested that I take the entrance exam for colleges.’

(15) a. e₁ > e₂ (matrix event follows complement event)

Zhangsan shuo/renwei Lisi shuo huáng.
Zhangsan say/think Lisi tell lie

‘Zhangsan said/thinks Lisi told lies.’

b. e₁ ∩ e₂ (matrix event overlaps with complement event)

Zhangsan shuo/renwei Lisi zài xízào.
Zhangsan say/think Lisi PROG take-a-bath

‘Zhangsan said/thinks Lisi was/is taking a bath.’

c. e₁ < e₂ (matrix event precedes complement event)

Zhangsan shuo/renwei Lisi huì chūlì.
Zhangsan say/think Lisi will handle
I agree with [?] that this restriction of temporal reference can be explained by the semantics of the control predicate, which will be discussed in next section.

3 Background Assumptions on the Semantics

3.1 Restructuring Control Predicates

Given that the predicates such as *qing* ‘invite’ and *bi* ‘force’ are partial control predicates, I adopt the analysis on partial control by [?], arguing that partial control is licensed by the temporal properties of the attitude verb. The complement of a control predicate denotes a property rather than a proposition. Take future-oriented *bi* ‘force’ as an example. (16) means for all buletic centred worlds (the world-time-individual triples) that are compatible with the fulfillment of attitude holder x’s desires in w at t for u to be y in w’ and it is compatible with x’s beliefs in w at t for t to be t’, there is another centred world ⟨w”, t”, z⟩ such that P holds for y to be z in w” and for t’ to be t” such that t’ precedes t”.

(16) \[ [\text{bi}]^{c,g} = \lambda P \lambda t \text{Top} \lambda t_0 \exists e \exists w[P(e)|w] \wedge t_{\text{Top}} \wedge t_0 < t_{\text{ana}} \subseteq \text{Rstate}(e, P) \]

Where for any pair of world-time-individual triples ⟨w, t, x⟩ and ⟨w’, t’, y⟩, ⟨w’, t’, y⟩ is a future-oriented extension of ⟨w, t, x⟩ iff for every α, β such that α is a coordinate of ⟨w, t, x⟩ and β is a coordinate of ⟨w’, t’, y⟩ of the same type as α, and α < _precedes_ β.

3.2 The Semantics for Perfective Aspect Markers

We assume that verbs and adjectives have an eventuality argument e and the run time of an event is obtained through the trace function τ, namely, the run time of e is τ(e). Complex eventualities denoted by telic sentences consist of a process eventuality (Inner stage) and a result state eventuality (Rstate) or target state eventuality.

Following [?], [?], we define the meaning of *le* and *guo* as below:

(17) \[ [\text{le}] = \lambda P \lambda t_{\text{Top}} \lambda t_0 \exists e \exists w[P(e)|w] \wedge \tau(\text{Istage}(e, P)) \subseteq t_{\text{Top}} \wedge t_{\text{ana}} \subseteq \text{Rstate}(e, P) \]

(18) \[ [\text{guo}] = \lambda P \lambda t_{\text{Top}} \lambda t_0 \exists e \exists w[P(e)|w] \wedge \tau(\text{Istage}(e, P)) \subseteq t_{\text{Top}} \wedge t_{\text{ana}} < t_0 \]

The semantics in (??) means that *le* requires the inner stage of an event denoted by P to be included within the topic time t_{Top} which in turn precedes the evaluation time t_0. The time of the result state includes a time t_{ana}, an anaphor-like variable that needs to be bound or given a value from context.

The semantics in (??) means *guo* requires that the time of the inner stage of an eventuality is included within the topic time t_{Top}, which in turn precedes the
evaluation time $t_0$. Unlike *le, guo also has a presupposition to account for the Discontinuity Effect. Since the presupposition of *guo does not contribute to the key of our analysis, we will not go further about it in later discussion, see [?] for an analysis.

4 Deriving Actuality Entailments

In this section, we will present our analysis with sentences in (5), repeated below as an example.

(19) a. Zhangsan bi Lisi chi guo fan, #keshi ta mei chi.
    Zhangsan force Lisi eat EXP rice but he NEG eat
    ‘Zhangsan had forced Lisi to eat, #but Lisi didn’t eat.’

b. Zhangsan bi guo Lisi chi fan, keshi ta mei chi
    Zhangsan force EXP Lisi eat rice but he NEG eat
    ‘Zhangsan had forced Lisi to eat, but Lisi didn’t eat.’

The denotation of a sentence without an aspect marker is given in (20). To simplify, we abbreviate the sentence in (20) as $P$ in (21). With *guo on the matrix predicate *bi, the sentence in (21) means *guo($P$) is true in the actual world $w$ if and only if the run time of the event such that *Zhangsan forces Lisi to eat is wholly before the speech time $s^*$. Since *guo only requires the forcing event to occur in the actual world, the event of Lisi’s eating in Zhangsan’s desire worlds is not affected by the perfective. Hence whether there is an actual event of eating does not matter, as expected in (19b).

(20) $[Zhangsan bi Lisi chi fan] = \lambda w \forall (t', w'. Lisi) \in \text{Bul}_{Zhangsan, w, t}$
    \[\exists \langle w'', t'', Lisi \rangle \in \text{Bul}_{Zhangsan, w, t} (\langle w'', t'', Lisi \rangle \text{ is a future-oriented extension of } \langle w', t', Lisi \rangle) \land \text{eat}(Lisi)(t'')(w'')\]

(21) $[19b] = \exists e \exists w[P(e)(w) \land \tau(Istage(e, P)) < s^*]$

When the perfective aspect marker is embedded in (19a), the structure reduces its clause size further. Namely, TP gets truncated and *bi ‘force’ combines with a predicate of event (embedded vP), forming a complex vP. Semantically, via Intensional Functional Application [?], it returns a predicate of event in (22).

\begin{center}
\begin{tikzpicture}
    \node (asp) {Asp}
    \node (guo) [below=of asp] {guo} child {node (v) {vP}};
    \node (zhangsan) [below=of v] {Zhangsan} child {node (v_dash) {v'} child {node (bi) {bi Lisi, 'force Lisi'} child {node (vp) {vP}}}};
    \node (pro) [below=of bi] {PRO, chi-fan 'eat'};
\end{tikzpicture}
\end{center}
Semantically, (23) means for all possible worlds that are compatible with Zhangsan’s desires in the actual world at the topic time $t_{\text{Top}}$, there is a centred world such that Lisi eats at $t''$ later than $t'$. $t'$ is compatible with Zhangsan’s actual perception of $t_{\text{Top}}$ and $t''$ precedes the speech time $t_0$.

(23) \[
\left[19a\right] = \forall \langle w', t', \text{Lisi} \rangle \in \text{Bul}_{\text{Zhangsan}, w_0, t_{\text{Top}}} \exists \langle w'', t'', \text{Lisi} \rangle \\
\langle w'', t'', \text{Lisi} \rangle \text{ is a future-oriented extension of } \langle w', t', \text{Lisi} \rangle \\
\text{and there exists an event } e \text{ such that } \text{eat}(e)(w'') \wedge \tau(I_{\text{stage}}(e, \text{eat}) \subseteq t'' \wedge t'' < t_0)\]

Now we get an actual event of Zhangsan forcing Lisi and in all of Zhangsan’s desire worlds, there is a past event of Lisi eating, with a slightly different meaning from (19b). Via the Event Identification across Worlds Principle (PEIW) suggested by Hacquard in (??), we further obtain that this actual event is Lisi’s eating. Hence the actuality entailment of (19a) is derived.

(24) **Event Identification across Worlds**: For any $w_1$, $w_2$, if an event $e$ occurs in $w_1$ and $w_2$, and $e$ is a P-event in $w_1$, it is a P-event in $w_2$ as well.

5 Conclusions

In this paper, I have focused on actuality entailment in MOC predicates taking a TP complement clause. Actuality entailment occurs when perfective aspect markers are embedded in the complement clause of a restructuring control predicate in Mandarin. In these restructuring cases, the matrix predicate and the embedded predicate form one single event. Via the PEIW, actuality entailment is derived when the perfective aspect scopes over the single event.

References

Semantically, (23) means for all possible worlds that are compatible with
Zhangsan's desires in the actual world at the topic time \( t_{\text{Top}} \), there is a centred
world such that Lisi eats at \( t' \) later than \( t' \). \( t' \) is compatible with Zhangsan's
actual perception of \( t_{\text{Top}} \) and \( t'' \) precedes the speech time \( t_0 \).

(23) \[ \llbracket 19a \rrbracket = \forall \langle w', t', Lisi \rangle \in Bul_{Zhangsan,w_0,t_{\text{Top}}} \exists \langle w'', t'', Lisi \rangle \]
\[ \langle w'', t'', Lisi \rangle \text{ is a future-oriented extension of } \langle w', t', Lisi \rangle \] & \[ \exists e \left[ \text{eat}(e)(w'') \land \tau[Istage(e,eat) \subseteq t''] \land t'' < t_0] \right] \]

Now we get an actual event of Zhangsan forcing Lisi and in all of Zhangsan's
desire worlds, there is a past event of Lisi eating, with a slightly different mean-
ing from (19b). Via the Event Identification across Worlds Principle (PEIW)
suggested by Hacquard in (??), we further obtain that this actual event is Lisi's
eating. Hence the actuality entailment of (19a) is derived.

5 Conclusions
In this paper, I have focused on actuality entailment in MOC predicates taking
a TP complement clause. Actuality entailment occurs when perfective aspect
markers are embedded in the complement clause of a restructuring control pred-
icate in Mandarin. In these restructuring cases, the matrix predicate and the
embedded predicate form one single event. Via the PEIW, actuality entailment
is derived when the perfective aspect scopes over the single event.

References
Proceedings of WC-
CFL. Vol.17. (1999)
2. Ernst, Thomas, Negation in Mandarin Chinese. 
Natural Language and Linguistic
3. Grano, Thomas, 
Control and Restructuring , Oxford University Press (2015)
4. Grohmann, Kleanthes, 
Prolific domains: On the anti-locality of movement depen-
5. Hacquard, Valentine, Restructuring and implicative properties of volere. 
Proceedings
6. Hu, Jianhua, Haihua Pan, and Liejiong Xu, Is there a finite vs. nonfinite distinction
7. Huang, C.-T. James, Pro-drop in Chinese: A generalized control theory. In:
The null
ciation with focus. 
Linguistics 39: 703-731(2001)
9. Lin, Jo-Wang, Temporal reference in Mandarin Chinese, 
Journal of East Asian
10. Lin, Jo-Wang, Aspectual selection and negation in Mandarin Chinese. Linguistics
12. Lin, Jo-Wang, Predicate restriction, discontinuity property and the meaning of
the perfective marker guo in Mandarin Chinese, Journal of East Asian Linguistics :
237-257 (2007)
13. Pearson, Hazel The semantics of partial control, Nat Lang Linguist Theory, 
14. Wurmbrand, Susi Restructuring cross-linguistically Proceedings of the North East-
15. Xiang, Yimei, Obligatory neg-raising in Mandarin: Negatives and aspects (2014)
16. Xu, Liejiong, Towards a lexical-thematic theory of control. The Linguistics Review
5:345-376 (1985)
17. Zhang, Niina Ning, Identifying Chinese dependent clauses in the forms of subjects. 
Optional negative concord with Turkish neither..nor

Paloma Jeretić
New York University

Abstract. Turkish is analyzed a strict negative concord language, where negative elements must occur with sentential negation. Turkish neither..nor phrases stand out of this system: sentential negation appears only optionally. In this paper I show that despite appearances, Turkish neither..nor phrases are objects similar to other n-words; this optionality is simply explained by a structural ambiguity, based on whether the coordinated constituents are of a propositional type.

1 Introduction

1.1 Strict negative concord in Turkish

Turkish has been analyzed as a strict negative concord language by [6], [12], [13], where negative elements cannot appear without sentential negation. An example of this phenomenon is seen below.

(1) Hiç-kimse gel-* (me)-di.  
    ever-anyone come-* (NEG)-PAST  
    No-one came. *No-one didn’t come.

The quantifier hic-kimse (no-one) cannot appear in a sentence without there being a form of sentential negation. hic-kimse is an n-word, according to definition (2).

(2) An n-word [3] is an element that (a) can be used in structures containing sentential negation or another a n-word yielding a reading equivalent to one logical negation, (b) can provide a negative fragment answer.

Using this term, we can define the meaning of a strict negative concord language:

(3) A strict negative concord language [3] is one in which all “n-words” co-occur with sentential negation.
1.2 Ne..ne: an unsystematic n-word.

Ne..ne (neither..nor) phrases appear to challenge the Turkish negative concord system, by only optionally co-occurring with sentential negation:

(4) Ne Ali ne (de) Beste gel-(me)-di.
    NE Ali NE (also) Beste come-(NEG)-PAST
    Neither Ali nor Beste came.¹

Ne..ne phrases are indeed n-words according to definition (2), because they can co-occur in structures containing a sentential negation (eg. (4)), and they can be used as fragment answers, as shown in the following example.

    who came   NE Ali NE Beste

This challenges the idea that Turkish is a strict negative concord language, or the definition of strict negative concord. This paper provides an explanation for what kind of object ne..ne is and its unexpected behavior.

1.3 Departing from Şener and İssever 2003

This phenomenon has been noticed by [2], [4], among others, and by [10], who provide an analysis for the phenomenon. They observe that if a ne..ne phrase is focused, then the predicate must be affirmative, and when the predicate is negative, the ne..ne phrase cannot be focused. Their analysis relies on the interaction between focal and negative features, in which the semantic negation of the Neg head is realized only when both of these features are checked.

The present paper argues against this analysis, by presenting additional facts and an updated generalization. Furthermore, Şener and İssever’s analysis fails to account for the difference between ne..ne phrases and other n-words. In this paper, I account for the observed behavioral difference, approaching the problem from the vantage point of the negative concord literature.

1.4 Goal and proposal

The negative concord literature does not discuss the phenomenon presented in this paper, in which a particular object, like Turkish ne..ne, stands out of the negative concord system of a language. My goal is to account for the optionality of negative concord in the presence of a ne..ne phrase, through which I provide support for and enrich the existing understanding of negative concord.

I argue that the difference between sentences in which ne..ne phrases don’t co-occur with sentential negation and those which do is structural. I provide

¹ The particle de is always optional, in sentences with and without negative concord. I allow myself to disregard the contribution of this particle in the rest of the paper.
evidence for the fact that the presence of sentential negation depends on whether the coordinated constituents are of a propositional type. If they are, then there is no negative concord reading. If they aren’t, then there is. I portray this below.

Example (6) has a ne..ne phrase that appears without sentential negation. I argue that its coordination structure can only be clausal, as shown in (7).

(6) Ne Ali ne Beste gel-di.
   ne Ali ne Beste come-past.
   Neither Ali nor Beste came.

(7) Structure for (6):
   a. ✓ [[Ne [Ali geldi]] [ne [Beste geldi]]]
   b. * [[[Ne Ali] [ne Beste]] geldi]

Example (8) has a ne..ne phrase that co-occurs with sentential negation. Only the structure in (9a) is available, where the coordinated constituents are not clauses. (9b), where constituents are clauses, only has a double negation reading, i.e. there is no negative concord reading available.

(8) Ne Ali ne Beste gel-me-di.
   ne Ali ne Beste come-neg-past.
   Neither Ali nor Beste came.

(9) Structure for (8):
   a. ✓ [[[Ne Ali] [ne Beste]] gelmedi]
   b. * [[[Ne [Ali gelmedi]] [ne [Beste gelmedi]]]

Once this generalization is posited, I propose a semantics of ne..ne as a non-negative existential quantifying over the domain spelled out by the members of the coordination. Further, I use Zeijlstra’s [12],[13] syntactic agreement theory of negative concord, and assume the following:

1. Uninterpretable negative features [uNeg] are carried by: sentential negation and all n-words, including ne..ne phrases.
2. These uninterpretable negative features invoke a null negative operator Op¬, that is semantically negative and carries interpretable negative features [iNeg].
3. This invocation happens locally, at the moment the [uNeg] carrier is merged.
4. Op¬ is of type <st,st>. If, at the moment of its invocation, there is a type mismatch, Op¬ does not enter the derivation.

These assumptions, together with the observed syntactic ambiguity, are sufficient to explain the full picture of Turkish negative concord.

2 Empirical generalization: a structural difference

This section presents evidence for the structural difference between ne..ne sentences without sentential negation (abbreviated as NE-Ø) and ne..ne sentences
with sentential negation that have a reading equivalent to one logical negation, i.e. engaging in negative concord (abbreviated as NE-NC).

A note on terminology: In these coordination structures, there is contrastive and non-contrastive material. Contrastive material is different in each coordinated constituent; non-contrastive material is identical in each constituent, each instance of non-contrastive material is a copy.

2.1 Coordinating clauses: only NE-∅
In this section, I show that NE-∅ sentences can coordinate clauses, and that NE-NC cannot.

Overt clausal coordination. The following example shows a NE-∅ sentence overtly coordinating a clause.

(10) Ne Ali dans et-ti ne Beste şarkı söyle-di.
    Ne Ali dance do-past ne Beste song say-past
    Neither Ali danced nor Beste sang.

If the verbs in the coordinated constituents are negated, only a double negation reading is available, i.e. there is no negative concord reading:

(11) Ne Ali dans et-me-di ne Beste şarkı söyle-me-di.
    Ne Ali dance do-neg-past ne Beste song say-neg-past
    Neither Ali didn’t dance nor Beste didn’t sing.

Verbal coordination. Any constituent overtly inside a ne..ne coordination that contains the verb will display the same pattern with clausal coordination. Below is an example of verbal coordination.

(12) Ali-yi ne gör-dü-m ne duy-du-m.
    Ali-acc ne see-past-1sg ne hear-past-1sg
    I neither saw nor heard Ali.

If sentential negation is expressed, only a double negation reading is available.2

(13) Ali-yi ne gör-me-di-m ne duy-ma-di-m.
    Ali-acc ne see-neg-past-1sg ne hear-neg-past-1sg
    I neither didn’t see nor didn’t hear Ali. (=I both saw and heard Ali)

This section shows that ne..ne can coordinate clauses and verbs, but if it does, it will not engage in negative concord.

2 In Turkish, verbal morphemes, including the sentential negation -mE, must be adjacent to the verb they attach to, i.e. cannot select coordinated vPs.

(1) *Ali ne gel ne ara-(ma)-di.
    Ali ne come ne call-(neg)-past.
2.2 Elliptical structures

We can recover typical elliptical structures in which a copy of a non-constrastive verb is not repeated in the second constituent, as in the following example.

(14) Ne Ali geldi, ne Beste.
     ne Ali came ne Beste
     Neither Ali came, nor did Beste.

This is a typical elliptical structure that involves clausal coordination, whose constituency would look like this: [Ne Ali geldi] [ne Beste geldi].

This evidence suggests that ellipsis is possible in general in ne..ne phrases. Thus, a typical example like (15) may have the following constituency structure: [Ne Ali geldi] [ne Beste geldi].

(15) Ne Ali ne Beste geldi.
     ne Ali ne Beste came
     Neither Ali nor Beste came.

In contrast, when we try to negate the predicate when it appears in the first constituent, the only reading available is a double negation reading, as predicted by the previous section:

(16) Ne Ali gel-me-di, ne Beste.
     ne Ali come-neg-past ne Beste
     Neither Ali didn’t come, nor did Beste. (=both came)

This suggests the possibility of verbal ellipsis in NE-∅ sentences, and the incompatibility of verbal ellipsis with NE-NC sentences. These facts are confirmed by evidence from gapping structures.

Gapping structures. There are certain constructions that force an ellipsis analysis, in particular, gapping structures ([9], a.m.o.) A gapping structure is the apparent coordination of non-constituents, e.g. when two constituents are arguments or adjuncts of a verb that is not overtly realized. We show below that NE-∅ is compatible with such a construction, but not NE-NC.

(17) Ne Ali pazartesi günü ne Beste sabah günü gel-di.
    Ne Ali monday day ne Beste tuesday day come-past
    Neither Ali came on Monday nor did Beste on Tuesday.

(18) ?/# Ne Ali pazartesi günü ne Beste sabah günü gel-me-di.
    Ne Ali monday day ne Beste tuesday day come-neg-past
    only possible reading: Neither Ali didn’t come on Monday neither Beste didn’t come on Tuesday. = Ali came on Monday and Beste came on Tuesday.3

3 This reading is attested only by some speakers, and is always difficult. This is no surprise, given all the work that must be done to get to it: ellipsis of something that is usually not under ellipsis; double negation twice (including once where one of the negations is elided).
This is evidence that NE-∅ can coordinate clauses in general, and that NE-NC is incompatible with clausal coordination.

2.3 NE-∅ is always clausal coordination

*Evidence from focus.* As noticed by [10], prosodic focus on non-contrastive constituents is incompatible with NE-∅.

(19) *Ne Ali ne Beste [gel-di]_F.
    Ne Ali ne Beste come-past

(20) *Ne Ali ne Beste [ev-e]_F  gel-di.
    Ne Ali ne Beste [house-dat]_F come-past

It is, in contrast, compatible with NE-NC.

(21) Ne Ali ne Beste [gel-me-di]_F.
    Ne Ali ne Beste come-neg-past
    Neither Ali nor Beste came.

(22) Ne Ali ne Beste [ev-e]_F  gel-me-di.
    Ne Ali ne Beste [house-dat]_F come-neg-past
    Neither Ali nor Beste came home.

Focused constituents cannot be elided. If NE-∅ sentences were obligatorily clausal coordinations, they would disallow focus on non-contrastive constituents. This is what we observe. On the other hand, NE-NC sentences do not coordinate clauses. And correspondingly, elements outside the coordination are able to be focused.

*Evidence from right dislocation.* In Turkish, backgrounded elements can be right-dislocated (see [11], [7], [8], [5], a.o.) Only full constituents can undergo movement. In NE-NC sentences, ne..ne phrases are constituents. And indeed, they can be left-dislocated, as shown below.

(23) Gel-me-di,  ne Ali ne Beste.
    come-neg-past, ne Ali ne Beste

In NE-∅ sentences, ne..ne phrases are not constituents. And correspondingly, they cannot be left-dislocated.

(24) *Gel-di,  ne Ali ne Beste.
    come-past, ne Ali ne Beste

(Data noticed by [10])

2.4 Interim conclusion

This section has shown that NE-∅ sentences must involve overt or covert verbal coordination, and that NE-NC cannot. The rest of the paper assumes clausal coordination for NE-∅ sentences, and coordination of constituents not including a verb for NE-NC sentences.
3 Analysis

In this section I provide an analysis of ne..ne that unifies it with other n-words, under Zeijlstra’s ([12], [13]) analysis of them as semantically non-negative existential quantifiers that carry uninterpretable negative features. The structural difference observed in the previous section will then explain the optionality of negative concord with ne..ne.

3.1 Semantics and negative features of ne..ne phrases

I propose that ne..ne phrases are formed by combining an operator NE with a tuple of arguments of same type. NE is an existential quantifier, that bears uninterpretable negative features [uNeg], in the spirit of Zeijlstra [12], [13].

Semantics of NE: an existential quantifier. The semantic value of NE corresponds to an operator that selects an n-tuple of arguments of the same type and existentially quantifies over that set of arguments. It doesn’t have any negative component.

\[
\text{NE} = \lambda (\alpha_1, \ldots, \alpha_n). \lambda \beta. \exists a \in \{\alpha_1, \ldots, \alpha_n\}. \beta(a) \text{ or } a(\beta)
\]

where \(\langle \alpha_1, \ldots, \alpha_n\rangle\) is an n-tuple of elements each of the same conjoinable type \(<x,y>\), and \(\beta\) is of type \(x\) or \(<x,y,z>\) (i.e. \(\beta\) selects or is selected by \(\alpha_i\)).

The ne particles therefore mark the left edges of each element of the set that the existential quantifies over; they don’t have a syntactic or semantic meaning of their own.

Syntactic features of NE: [uNeg].

Zeijlstra’s theory of negative concord. Zeijlstra [12], [13] takes negative concord to be a syntactic Agree relation (following minimalist ideas on agreement, cf. [1]) between a single interpretable negative feature [iNeg] and one or more uninterpretable negative features [uNeg]. In strict negative concord languages, n-words are analyzed as existential quantifiers carrying [uNeg], and sentential negation also carries [uNeg]. These uninterpretable negative features must be checked off by a covert negative operator Op\(\neg\) that has interpretable negative features [iNeg]. Semantic negation is only interpreted once, where Op\(\neg\) is located.

Zeijlstra’s proposal applied to example (1):

\[
\text{Op}\neg \quad \text{hičimse gel - me - di}
\]

\[
\text{iNeg} \quad \text{[uNeg]} \quad [\text{uNeg}]
\]

Ne..ne analyzed as a typical n-word. Despite appearances, I take ne..ne phrases to be similar objects to other n-words, and analyzed in the same way under Zeijlstra’s theory of negative concord. This means that while the semantics of NE doesn’t have a negative component, syntactically, NE carries uninterpretable features [uNeg], like other n-words, that must agree with a clausemate [iNeg].
3.2 Assumptions

Following Zeijlstra’s [12], [13] syntactic agreement theory of negative concord, I make the following assumptions:

A1. Uninterpretable negative features [uNeg] are carried by: sentential negation and the existential quantifiers that form n-words, including those of ne..ne phrases;

A2. [uNeg] triggers the appearance of a null semantically negative operator Op¬ that carries interpretable negative features [iNeg];

A3. The appearance of Op¬ happens locally, at the moment a [uNeg] carrier is merged, and only then;

A4. Op¬ is of type <st,st>, with denotation λp_s,t.λw.¬p_w. If, at the level the [uNeg] carrier is merged, there is a type mismatch with the constituent to be selected, Op¬ does not enter the derivation.

3.3 Explaining the basic negative concord facts in Turkish

The explanation in a nutshell. There are two cases:

1. If a [uNeg]-carrying object is merged at a level that yields a propositional type <s,t>, Op¬ may merge. The only such object is a ne merging with a tuple of clauses. This is the case of a ne..ne phrase not co-occurring with sentential negation.

2. If a [uNeg]-carrying object is merged at a level that does not yield an object of a propositional type <s,t>, Op¬ may not merge. However, the [uNeg] needs to be checked off, which means that another [uNeg]-carrying object needs to merge higher to trigger the appearance of Op¬, which can be the job of a sentential negation marker. The objects relevant to these cases are existentials that form n-words like hickimse, or a ne that combines with objects smaller than propositions. Because of the necessity of sentential negation, these are cases of negative concord.

Accounting for NE-Ø sentences. Consider the typical NE-Ø sentence below.

(27) Ne Ali ne Beste geldi.
    ne Ali ne Beste come-past.
    Neither Ali nor Beste came.

I take this sentence to be formed by adjoining NE to the tuple of TPs ⟨Ali geldi, Beste geldi⟩. The semantic derivation looks like the following:

(28) [[NE][[Ali geldi, Beste geldi]]] = λw.∃p_{<s,t>} ∈ {Ali geldi, Beste geldi},p_w
     = λw.∃T ∈ {Ali geldi,w, Beste geldi,w}.T = 1

Since NE carries uninterpretable negative features (by A1), the abstract operator is invoked at the moment it is merged (by A2 and A3). Since the current derivation is of type <s,t>, the operator Op¬, of type <st,st>, can apply (by A4). The result is shown here.
(29) \(\operatorname{Op}^{-}[28] = \lambda p_{<s,t>} \cdot \lambda w. \neg p_w [\lambda w. \exists T \in \{\text{Ali geldi}_w, \text{Beste geldi}_w\}] = \lambda w. \neg \exists T \in \{\text{Ali geldi}_w, \text{Beste geldi}_w\}. T = 1\)

A pictorial representation of the derivation, bringing semantics and syntactic features together, can be seen below.

(30) \(\text{Ne Ali ne Beste geldi.}\)

\[
\begin{array}{c}
\text{TP} \\
\lambda w. \neg \exists T \in \{\text{Ali geldi}_w, \text{Beste geldi}_w\}. T = 1
\end{array}
\]

\[
\begin{array}{c}
\text{[uNeg]} \\
\lambda w_{\neg} \cdot \lambda w. \neg p_w \\
\text{TP} \\
\lambda p_{\neg} \cdot \lambda w. \neg p_w \\
\text{[INeg]} \\
\lambda w. \exists T \in \{\text{Ali geldi}_w, \text{Beste geldi}_w\}. T = 1
\end{array}
\]

We end up with the desired interpretation of (27).

**Accounting for NE-NC sentences.** When \(\text{Ne}\) adjoins to a tuple of elements that don’t have a propositional type, \(\operatorname{Op}^{-}\) is invoked, but because of its type restriction \(<st, st>\), it will not be able to appear. It will only appear later if another [uNeg] enters the derivation and makes it of type \(<s, t>\). Consider the typical NE-NC sentence below.

(31) \(\text{Ne Ali ne Beste gel-me-di.}\)

\(\text{ne Ali ne Beste come-neg-past.}\)

Neither Ali nor Beste came.

In this case, I claim that \(\text{ne}\) selects for the tuple of DPs \(\text{Ali}\) and \(\text{Beste}\), and the syntactic derivation looks like this.

(32) \(\text{[[[ [\text{ne} (\langle \lambda.Q.Q(\text{Ali}), \lambda.Q.Q(\text{Beste}) \rangle) gel] -me] -di]}\]}

The first step of the derivation involves \(\text{ne}\) merging with the tuple of generalized quantifiers \(\langle \lambda.Q.Q(\text{Ali}), \lambda.Q.Q(\text{Beste}) \rangle\). This can be seen below.

(33) \(\text{NE}((\lambda.Q.Q(\text{Ali}), \lambda.Q.Q(\text{Beste}))) = \lambda P. \lambda w. \exists x \in \{\lambda.Q.Q(\text{Ali}), \lambda.Q.Q(\text{Beste})\}. x(P_w)\)

The negative operator \(\operatorname{Op}^{-}\) is invoked due to the [uNeg] of \(\text{Ne}\), however, there is a type mismatch, so \(\operatorname{Op}^{-}\) does not enter the derivation. Next, the verb \(\text{gel}\) is merged:

(34) \(\lambda w. \exists x \in \{\lambda.Q.Q(\text{Ali}), \lambda.Q.Q(\text{Beste})\}. x(\text{gel}_w) = \lambda w. \exists T \in \{\text{gel}_w(\text{Ali}), \text{gel}_w(\text{Beste})\}. T = 1\)
The uninterpretable features of \( \text{NE} \) are waiting to be checked off. The only way of doing this is by merging some other \([\text{uNeg}]\)-carrying object. This is the job of a sentential negation, whose denotation is the identity function on intensional propositions.

\[(35) \quad [-\text{mE}][(34)] = \lambda p_w. \lambda w. p_w(\lambda w. \exists T \in \{\text{gel}_w(\text{Ali}), \text{gel}_w(\text{Beste})\}.T = 1) = \lambda w. \exists T \in \{\text{gel}_w(\text{Ali}), \text{gel}_w(\text{Beste})\}.T = 1\]

Finally, the \([\text{uNeg}]\) of the sentential negation \([-mE]\) invokes \(\text{Op}^-\), now able to merge:

\[(36) \quad \lambda p_w. \lambda w. \neg p_w(\lambda w. \exists T \in \{\text{gel}_w(\text{Ali}), \text{gel}_w(\text{Beste})\}.T = 1) = \lambda w. \neg \exists T \in \{\text{gel}_w(\text{Ali}), \text{gel}_w(\text{Beste})\}.T = 1\]

A pictorial representation of the derivation, bringing semantics and syntactic features together, can be seen below.

\[(37) \quad \text{Ne Ali ne Beste gelmedi}.
\]

We end up with the desired interpretation of (31).

*Negative concord with other n-words.* The process is essentially the same as the one for non-propositional \(\text{ne..ne}\) phrases, described above. Zeijlstra’s proposal hinges on the fact that quantifiers like “nobody” carry uninterpretable
negative features. I introduce more precision by saying that there is an existential quantifier, call it $\exists_{\neg\text{Neg}}$, with denotation $\lambda P.\exists x P(x)$, that carries uninterpretable features $\text{Neg}$. This quantifier may merge with a property of type $<e,t>$, e.g. “person”, “thing”, “place”, “time”, yielding “nobody” (hiç kimse), “nothing” (hiçbir şey), “nowhere” (hiç bir yer), “never” (hiç bir zaman; asla).

At the moment the existential quantifier merges with a property, the appearance of $\text{Op}\neg$ is triggered. It will try to apply to the relevant generalized quantifier, e.g. $\lambda P.\lambda w.\exists x. P_w(x) \& \text{person}_w(x)$, but will fail due to a type mismatch. The derivation then goes on exactly in the same way as non-propositional ne..ne: the only way of checking off the $\text{Neg}$ of $\exists_{\neg\text{Neg}}$ is to merge another $\text{Neg}$-carrying object that will invoke $\text{Op}\neg$ at a propositional level. This yields a sentence with negative concord. There is no non-negative concord counterpart, because $\exists_{\neg\text{Neg}}$ cannot merge with propositions.

This analysis gives an explanation for the necessity of the sentential negation marker with n-words.

4 Conclusion and discussion

This paper addresses the puzzle raised by the exceptional behavior of Turkish ne..ne phrases, in which they optionally (instead of obligatorily) give rise to sentential negation. This gives them the appearance of being a distinct object from the rest of the n-words.

However, I show in this paper that they are in fact a typical n-word. Their particular behavior is explained by their ability to select for clauses, which typical existential quantifiers don’t have. When ne..ne phrases coordinate propositional constituents, the null negation operator may merge, and therefore overt negative concord does not occur. When they coordinate non-propositional constituents, the negative operator cannot merge and the $\text{Neg}$ is left unchecked; another $\text{Neg}$ carrier must be merged, which is the job of a propositional ne..ne phrase or a sentential negation marker.

The success of this analysis of an unobserved puzzle for negative concord gives support to Zeijlstra’s theory, in that it includes ne..ne phrases in the set of $\text{Neg}$ carriers, that all have the same behavior, and only differ in their type. In addition, it addresses some problems that his system had: Zeijlstra is unable to explain the necessity of sentential negation when other n-words are present; this paper provides an explanation for it.

While this paper is local to Turkish, it gives rise to questions about negative concord in general. My fieldwork shows that there are languages in which “neither..nor” phrase equivalents behave like n-words with respect to negative concord (eg. Russian, Spanish, French, Hungarian), other languages in which they don’t (Cairene Arabic, Farsi, Greek, Hebrew, Kurdish, Turkish), but no languages in which an n-word other than a “neither..nor” phrase stands out of the system. The obvious difference between “neither..nor” phrases and typical quantifiers is the fact that the former uses coordination structures, while the lat-
ter don’t. A structural explanation for the difference between “neither..nor” and other quantifiers is thus in order, as proposed in this paper. If intra-linguistic variation in negative concord is caused by the structure, could cross-linguistic variation in negative concord be explained similarly? This is a question left for further research.

References

10. Şener, S., İşevey, S.: The interaction of negation with focus: ne... ne... phrases in turkish. Lingua 113(11), 1089–1117 (2003)
An Analysis of Counteridenticals in Terms of Dream Reports

Carina Kauf
University of Göttingen
carina.kauf@uni-goettingen.de

Abstract This paper investigates the semantics of “counteridenticals” (Goodman 1984) and argues that they are best analyzed along the lines of dream reports. To this end, I show that counteridenticals and dream reports exhibit striking grammatical and perceptual parallels. I then propose an analysis of counteridentical meaning that constitutes a version of Percus and Sauerland’s (2003) dream report analysis modified by the notion of asymmetric be as proposed in Percus and Sharvit (2014).

Counteridenticals are counterfactual conditionals whose antecedent clauses identify two inherently incompatible entities with each other, e.g.

(1) If I were you, I’d buy the blue dress.
(2) If Peter were Angela Merkel, he’d be the chancellor of Germany.

For my analysis, I focus on counteridenticals of the form If x were y, x would VP, thus excluding structures like If I were you, your name would be Carina, which are marginal for many speakers.

What seems to be the case for the conditionals relevant to this paper is that the speaker imagines a counterfactual world in which the subject of the antecedent clause is identified with the clause’s object, creating a counterpart of the subject whose properties have been modified by those of another individual (i.e. the object). The consequent proposition is then evaluated with respect to this composed entity, for which it remains to be determined to what extent its properties were originally part of the object and to what extend they were part of the subject. In fact, the different property combinations of subject and object might predict two different readings: an ‘advice’ reading (exemplified by (1)) and an ‘imagine’ reading (exemplified by (2)).

Contrary to existing analyses of counteridentical meaning (Lakoff 1996; Kocurek 2016), I argue that counteridenticals are best analyzed along the lines of dream reports. To this end, I show that counteridenticals exhibit striking grammatical as well as perceptual parallels with dream reports. Some of these parallels have already been noted in the literature by Arregui (2007), and this paper provides two additional arguments in favor of an analysis which treats the two linguistic structures on a par. A novel approach is required since the existing proposals do not yield satisfactory solutions to at least two problems connected to counteridentical meaning:
1. How do we arrive at the composed individual that makes up the shared counterpart of the copular clause’s subject and object in the counterfactual worlds (e.g. the consequent pronouns I in (1) and he in (2))?

2. Once we have this individual, how can we get back to the distinct entities in line for reference (i.e. the composed, non-actual counterpart individual, and the two actual entities) by means of pronouns?

Existing proposals focus on providing a solution to the first problem and neglect the reference constraints counteridenticals impose on their consequent pronouns; hence, they make wrong predictions with respect to question 2. As to the composition of the counterpart individual, Lakoff’s theory is argued to undergenerate, whereas Kocurek’s proposal overgenerates.

The analysis proposed in this paper constitutes an adapted version of Percus and Sauerland’s (2003), henceforward P&S, account of dream reports which incorporates the notion of asymmetric be as proposed in Percus and Sharvit (2014). It not only considers the correlations between the two grammatical constructions but also gives answers to both of the questions stated above.

1 Parallels Between Counteridenticals and Dream Reports

Counteridenticals and dream reports exhibit at least five striking parallels with regard to their grammatical and perceptual make-up.

1. Both allow for sequences to occur that would not be allowed as independent matrix clauses.

   \begin{align*}
   (3) & \\
   & \text{a. } *I \text{ kiss(ed) me.} \\
   & \text{b. } I \text{ dreamed } I \text{ was Brigitte Bardot and } I \text{ kissed me.} \\
   & \text{c. } \text{If I were you, I'd kiss me.} \quad (\text{Arregui } 2007: 31)
   \end{align*}

   In extensional contexts (e.g. (3-a)) the sequence \textit{I kiss(ed) me} constitutes a violation of the binding principle B, which requires that a pronoun must be unbound within its governing category (cf. Chomsky 1981). In dream reports and counteridenticals, however, pronouns with the same features may have multiple referents, as can be made explicit by adding indices to the examples in (3), yielding (4). Whereas the subscripts \( i \) and \( j \) are used for pronouns referring to entities which inhabit the actual world in addition to the counterfactual one, the subscript \( i \oplus j \) designates pronouns referring to a non-actual entity composed of a combination of the antecedent clause’s subject’s and object’s properties, i.e. the subject’s dream/counterfactual self. The availability of multiple referents enables the circumvention of the binding principle’s application in (3)/(4).

   \begin{align*}
   (4) & \\
   & \text{a. } I_i \text{ dreamed } I_i \text{ was Brigitte Bardot and } I_{i\oplus j} \text{ kissed me}. \\
   & \text{b. } \text{If } I_i \text{ were } y_{i\oplus j}, \text{ } I_{i\oplus j} \text{ 'd kiss me}. 
   \end{align*}
2. Both allow principle B violations only for first person pronouns; the same structure is not permitted for third person pronouns (5-a), or second person pronouns (5-b) (cf. Arregui 2007: 31; see also Harley & Ritter 2002).

(5) a. (i) *If Peteri were Billj, hei⊕j'd kiss himi.
   (ii) *Suci dreamed [shei was Brigitte Bardotj and] shei⊕j kissed heri.

   b. (i) */?If youi were mej, youi⊕j'd kiss youi.
   (ii) *Youi dreamed [youi were Brigitte Bardotj], and youi⊕j kissed youj.

This observation is particularly striking considering that second and third person pronouns may actually have multiple referents in counteridenticals and dream reports (e.g. If Susani were Suej, shei⊕j would be in love with heri, brother). Note that I deliberately chose a sentence with a possessive as an example in order to avoid the intervention of the binding principle B. It serves to show that both the actual as well as the composed counterpart of the antecedent clause’s subject provide feasible referents for third person pronouns in the counterfactual world. The same observation also holds for second person pronouns.

3. Both enable us to comprehend clauses which, under canonical circumstances (i.e. excluding role playing situations, etc.), seem irremediably false in extensional contexts.

(6) a. *I was you.
   b. I dreamed I was youi.
   c. If I were youi, I would be happier. (cf. Arregui 2007: 31)

When evaluated against the facts of the actual world, the identification of two inherently different individuals seems clearly infelicitous. Nevertheless, in the cases of dream reports and counteridenticals, we can easily make sense of such a relation, since we derive that instead of consulting our knowledge of the actual world we are to imagine worlds which differ from ours with regard to some contextually relevant presuppositions, here: the identity of the speaker/the addressee.

4. The pronouns of both constructions obey the Oneiric Reference Constraint (ORC), a syntactic constraint on pronoun movement that rules out any LF for dream reports in which some pronoun referring to the dream-self is asymmetrically c-commanded by a pronoun referring to the actual entity (cf. P&S 2003: 5). The ORC explains why dream reports involving two pronouns with the same agreement features (e.g. (7)) are ambiguous between only three readings, even though there are four possible combinations of the consequent pronouns’ referents (i.e. the actual-John and his dream-self): It disallows that reading in which the first pronoun refers to the actual self of the dreamer, hect, while the second one refers to that person’s dream-self, hectj (cf. (7-d)).

(7) John dreamed that he was marrying his grand-daughter.

   a. In John’s dream, the hectj marries hisj grand-daughter.
   b. In John’s dream, hej marries hisi grand-daughter.
c. In John’s dream, he, marries his, grand-daughter.
d. *In John’s dream, he, marries the his, grand-daughter.

The pronouns in counteridenticals obey a similar constraint: Those pronouns which can be interpreted ambiguously between referring to the speaker’s actual self and the person s/he counterfactually identifies with obey the ORC (cf. (8)). (Note that the first consequent pronoun, Ii⊕j, is excluded from the constraint in this example since it can never refer back to the actual speaker). In (8), the ORC renders that reading infeasible, or at least marginal, in which the actual speaker’s son shall play with the counterfactually imagined daughter.

(8) If I were you, I’d want my son to play with my daughter.
  a. If Ii were youj, Ii⊕j’d want myi⊕j son to play with myj daughter.
  b. If Ii were youj, Ii⊕j’d want myi⊕j son to play with myj daughter.
  c. If Ii were youj, Ii⊕j’d want myi son to play with myj daughter.
  d. */? If Ii were youj, Ii⊕j’d want myi son to play with myi⊕j daughter.

5. Both counteridenticals and dream reports presuppose that the speaker has taken over the entire set of (contextually relevant) properties of the person s/he imagines to be. If the speaker wants to change any of the properties which undergo the re-ascription process, s/he has to make the change explicit.

(9) a. I dreamed I was you. But you lived in New York and had a great apartment. (Arregui 2007: 36)
   b. [Assuming Angela Merkel does not to like traveling.]
      If I were Angela Merkel, I’d be traveling all around the world, but (unlike her,) I’d love it.

The copular clauses raise certain expectations with regard to what it means for the dreamer to be the dreamee in the counterfactual worlds, or the subject of the counteridentical antecedent to be the object. If the speaker deviates from these expectations without explicating so in the utterance, the listener is expected to object:

(10) a. A: I dreamed I was you. I lived in New York and I had a great apartment...
    B: I don’t think it was me that you dreamed you were. My apartment is pretty crappy.
    b. A: If I were Angela Merkel, I’d be traveling all around the world and I’d be loving it.
    B: Wait a minute, I thought Angela Merkel hates traveling.

2 An Analysis of Counteridenticals in Terms of Dream Reports

Given the structural and conceptual parallels with dream reports, I propose an analysis of counteridenticals on a par with that of dream reports. Such a novel
approach is required since the few existing theories of If $x$ were $y$-conditionals are found to be problematic with respect to the composition of the counterpart individual and the ORC.

2.1 Discussion of Existing Analyses

Lakoff’s (1996) proposal for the meaning of counteridenticals is based on his assumption that each individual can be conceptualized as the combination of two separable parts: the individual’s Subject (i.e. one’s subjunctive/intrinsic properties) and his Self (i.e. one’s metaphysical situation). With regard to If $x$ were $y$-counteridenticals, this means that we have to take into account four distinct entities: Subject-of-$x$, Subject-of-$y$, Self-of-$x$, and Self-of-$y$.

Building on Lewis’ (1973) counterpart theory, Lakoff suggests that the meaning of If $x$ were $y$ comes about via a counterpart relation which creates a hybrid agent in the counterfactual antecedent-worlds that combines the Subject-of-$x$ from the actual world with the Self-of-$y$ from the actual world via identifying the two Subjects at hand with each other across worlds (i.e. in the counterfactual worlds: Subject-of-$x$ = Subject-of-$y$) (cf. Lakoff 1996: 93f). As a result, the antecedent of a counteridentical always creates a non-actual individual which possesses the addressee’s metaphysical and the speaker’s subjunctive properties, a strategy which Lakoff argues for on the basis of examples like the following:

(11)  
[Supposing I have done cruel things to you, but you are very forgiving. I, however, would not be forgiving in the same situation.]  
If I were you, I’d hate me.  
(= If I were in your situation, but had my own capacity for feelings, judgments, etc. instead of yours, I (as that hypothetical hybrid person) would hate me.)  
(Lakoff 1996: 94)

Such a theory undergenerates: Most notably, Lakoff’s strict rules for creating a counterpart predict the infeasibility of counteridenticals where the hybrid agent refers anaphorically to the antecedent clause’s object instead of the subject (cf. (12)), or where it is the metaphysical properties that are being transferred onto the subject and not the subjunctive ones (cf. (13)). Nevertheless, both of these examples are attested to be felicitous sentences of English in suitable contexts. Note that focus might play a role in these cases, as well (for a more detailed discussion of Lakoff’s theory, see Kauf (2016)).

(12)  
[While playing, Fred broke George and his mom’s favorite vase. Their mom is approaching and George asks Fred what he is going to do.]  
If I were you, I’d get upstairs this second! [You always run away from responsibility. I, however, will stay down here and explain the situation.]  
(13)  
If I were you, I’d be short and named George.

1 For an analysis in the same spirit, see Malamud (2006: 91–93).
The most recent as well as the most elaborate counterpart theory of counteridenticals to this day has been put forward by Kocurek (2016). As Lakoff’s theory, his proposal also constitutes a variation of Lewis’ (1973) counterpart theory. Nevertheless, Kocurek draws different conclusions from it. He exploits the fact that Lewis defines counterparthood on the basis of similarity, which leaves room for there being multiple counterparts of one single individual within the same world. These distinctive counterparts are derived via considering the individual under different notions of similarity/“senses” (Kocurek 2016: 23), creating correlates of the actual entity that are similar to it with regard to different aspects. Kocurek assumes with Lewis (1973) that the different underlying similarity relations can be made overt by means of finding appropriate modifier clauses for each term (cf. 210).

On the basis of these assumptions, Kocurek develops the following algorithm for analyzing the semantics of counteridenticals:

1. Regiment the English sentence without any counterpart indices.
2. Go term by term and determine what phrase of the form ‘as _’, ‘with _’, or ‘being _’ (or others like ‘qua _’, etc.) would be appropriate to modify that term with in that context.
3. Go back and assign counterpart indices to terms and counterpart relations to counterpart indices that would make sense of those phrases. Hereby, assign a separate index to each distinct way of filling in the blanks in 2.

(cf. Kocurek 2016: 24)

In practice, this algorithm works as follows (cf. ibid: 25):

(14) a. If I were you, I’d vote for me.
   b. \( (I^1 = you^2) > \text{VoteFor}(I^1, I^3) \)
   c. If I (with my beliefs and preferences) were you (as a voter), I (with my beliefs and preferences) would vote for me (as a candidate).

Since the theory predicts that only two terms which are counterparts of the same actual individual and which have been derived under the same similarity relation can be coreferential, Kocurek is able to account for the difference in meaning between (14) and (15).

(15) a. If I were you, I’d vote for myself.
   b. \( (I^1 = you^2) > \text{VoteFor}(I^1, I^1) \)
   c. If I (with my beliefs and preferences) were you (as a voter), I (with my beliefs and preferences) would vote for me (with my beliefs and preferences). (cf. ibid)

In contrast to Lakoff, Kocurek’s proposal overgenerates. One reason for this overgeneration is his treating the copula in the antecedent clause as an equative \((I^1 = you^2) > I^1 \ VP\). According to this theory, the antecedent clause creates a counterpart of the entities I and you which is I’s counterpart with respect to the the similarity relation 1 and you’s counterpart with respect to the similarity relation 2. Nevertheless, the equality relation between \(I^1\) and \(you^2\) in connection
with the Principle of Identity Substitution (Quine 1961: 113) then predicts the equivalence of the sentences in (16), even though they clearly differ in meaning.

\begin{alignat}{2}
&\text{a. If } I^1 \text{ were you}^2, I^1 \text{ would VP.} \\
&\quad (I^1 = you^2) > V(I^1) \\
&\text{b. *If } I^1 \text{ were you}^2, you^2 \text{ would VP.} \\
&\quad (I^1 = you^2) > V(you^2)
\end{alignat}

Furthermore, the disproportionate relation between the theory’s flexible rules for assigning modifying clauses to pronouns, and it’s being crucially dependent on assigning different counterpart indices to terms based on a difference in their modifying clauses, makes false predictions, as well.

Lastly, Kocurek’s basic assumption – the fact that the context may provide more than one counterpart for a given individual – can also be questioned. If it were the case that the availability of multiple counterpart relations for an individual is responsible for the difference between sentences (14) and (15), the same should also be possible for other individuals/pronouns apart from the speaker/I. Nevertheless, we have seen that such principle B violations are only feasible for first person pronouns (cf. section 1, similarities 1. and 2.). This motivates us to conclude with Arregui (2007) that each context only provides one counterpart for a given individual.

In addition to the discussed challenges for each of the counterpart theories, none of the existing proposals predict the connection to dream reports, or, more importantly, the ORC that counteridenticals impose on their consequent pronouns. Therefore, I propose a novel analysis of counteridentical meaning which takes these parallels into consideration instead of building on Lewis’s (1973) counterpart theory.

2.2 The Proposal

Recall that there are two problems which need to be solved with respect to counteridentical meaning: a) the compositional make-up of the non-actual hybrid agent, and b) the reference to the different entities relevant for the analysis by means of pronouns. Let us start by looking at the first question.

For a variety of reasons, most notably the meaning distinction between the antecedent clauses \textit{If I were you, [I would VP] and If you were me, [you would VP]}, which rules out an analysis that interprets both clauses by means of the relation $I = you$, as well as the overgeneration observed in Kocurek’s theory, I refrain from interpreting the counteridentical antecedent as an equative copular clause (for a more detailed discussion see Kauf 2016). Instead, I endorse the notion of asymmetric \textit{be} as proposed by Percus and Sharvit (2014), in its redefinition by Zhang (2016) (indicated below by $Z_{\text{h}}$), for my analysis.

Motivated by mistaken identity contexts, Percus and Sharvit propose the existence of an asymmetric copula $be_{\text{asymmetric}}$ which takes an individual concept (in its realization as a set of properties (cf. Zhang 2016)) as its input and identifies it with an individual $x$. If such an individual concept is overtly available, as in
Dan is the new PhD student, the concept (here: the new PhD student) is simply predicated as a property of the subject referent (here: Dan) by means of (17).

\[ \llbracket \text{PRED} \rrbracket^w_{\text{Zhang}} = \llbracket \text{be_asymmetric} \rrbracket^w_{\langle s, e, t \rangle} = \lambda y_c. \lambda x_c. P(w)(x) \]

For cases in which a person is identified with another person instead of an overt individual concept, Percus and Sharvit suggest to refine the semantics of asymmetric be such that one of the individuals, y, is first coerced into contextually salient set of properties before \( \llbracket \text{PRED} \rrbracket^w \) can be applied (cf. (18)).

\[ \llbracket \text{PRED} y \rrbracket^w_{\text{Zhang}} = \llbracket \text{be_asymmetric} \rrbracket^w_{\langle e, t \rangle} = \lambda y_c. \lambda x_c. P_{(w,y)}(w)(x), \]

where \( P_{(w,y)} \) of type \( < s, e, t > \) represents the coercion of the individual \( y \) into some contextually salient set of properties in a world \( w \).

Besides the conceptual appeal of an analysis of counteridentical antecedents as asymmetric copular clauses, adopting Percus and Sharvit’s notion of asymmetric be also enables us to account for the difference in meaning between the structures in (19-a) and (19-b).

\[ \begin{align*}
(19) & \quad \text{a. If I were you, [I would VP]} & \Rightarrow & \quad P_{(w,you)}(w)(I) \\
& \quad \text{b. If you were me, [you would VP]} & \Rightarrow & \quad P_{(w,i)}(w)(you)
\end{align*} \]

Furthermore, it allows us to account for the different degrees of identification between the antecedent clause’s subject and object that can be observed in counteridenticals, since it does not impose any restrictions on the set of properties which the object is coerced into. In (1), repeated for convenience in (20), the speaker assumes the addressee’s external properties while keeping his/her internal properties intact, a strategy which enables him/her to give advice. Note that in this case, the consequent property must neither be true of the subject nor the object in the actual world. By contrast, the truth of the consequent clause in (2)/(21) is achieved if Peter is either completely identified with Angela Merkel in the counterfactual worlds or if he is merely identified with her with respect to her profession. In this case, the consequent property is always true of the object in the actual world.

\[ \begin{align*}
(20) & \quad \text{If I were you, I’d buy the blue dress.} \\
(21) & \quad \text{If Peter were Angela Merkel, he’d be the chancellor of Germany.}
\end{align*} \]

The observed correlation between the distinct interpretations of the antecedent clause and the different combinations of subject and object properties might predict the the clause If \( x \) were \( y \) to be ambiguous between an ‘advice’ reading (cf. (20)) and an ‘imagine’ reading (cf. (21)). Interestingly, we indeed find evidence for such multiple readings cross-linguistically: At least Polish, LIBRAS and Greek all make a grammatical distinction between the antecedent clauses they allow for the two kinds of counteridenticals. In Polish, for example, the construction usually used to express counteridenticals, i.e. past tense-marking of the copula in combination with the subjunctive mood (\textit{Gdybym byl tob\d{a}}} [literally: I be-FAST you]), is restricted to the ‘imagine’ reading of counteridenti-
cals, even though the copula is not generally restricted to equative contexts. In
the case of ‘advice’ counteridentical, by contrast, speakers of Polish must make
use of a paraphrase structure (Na Twoim miejscu [= literally: On your spot])
p.c. Z. Fuchs; Kauf (2016)).
After having explicated the relation which I assume the counteridentical an-
tecedent clause to set up between its subject and object, let us now turn to the
question of how the distinct entities at stake (i.e. the composed, non-actual coun-
terpart individual and the two actual entities) can be referred to by means of
pronouns. In this context, it is especially the similarity with respect to the ORC
which calls for an analysis of counteridenticals along the lines of P&S (2003). In
their analysis of dream reports, P&S propose to make use of concept generators
in their realization as centered worlds:

\[
\text{[dream]}^g = \lambda P. \lambda x. \lambda w. \text{For all } <y, w'> \text{ in DREAM}_{x,w}, P(y)(w') = 1.
\]

(DREAM\(_{x,w}\) stands for the set of pairs \(<y, w'>\) such that \(w'\) is a world
compatible with \(x\)'s dream in \(w\), and \(y\) is the individual in \(w'\) who \(x\), in
\(w\), identifies as himself.) (ibid: 8)

Multiple pronoun reference is accounted for in the following way: Reference to
the actual person is realized by means of an unstarred pronoun, which is analyzed
in situ like a usual variable. It combines with a world parameter which, due to
lambda-abstraction, receives its denotation from the worlds compatible with the
agent’s dream worlds, i.e. \(w'\). Reference to the dream-self, on the other hand,
is realized via a starred pronoun, which behaves similar to a relative pronoun:
it does not receive an interpretation in situ but moves to the left periphery of
the complement clause, which triggers a predicate abstraction over the trace it
leaves behind (cf. P&S 2003: 7f). Since P&S assume the denotation of ‘dream’
to be similar to that of attitude verbs, i.e. they assume that ‘dream’ quantifies
over centered worlds and takes a property (the meaning of the complement
clause) as an input (cf. (22)), such a movement leads to an identification of the
moved pronoun with the center of worlds that are compatible with agent’s dream
worlds, i.e. the dream-self \(y\). A possible logical form of a dream report under this
proposal looks the following:

\[
\text{(John) dreamed that he}^{\text{dream-self}} \text{ was marrying his}^{\text{actual-self}} \text{ grand-daughter.}
\]

\[
a. \text{dream [ he}^{*} \lambda x. \lambda w. \forall <y, w'> \text{ in DREAM}_{x,w}, \ y \text{ marries the grand-daughter of}
g(2)(w') \text{ in } w'.
\]

\[
b. \text{This “property” will hold, e.g., of John, if he has a dream in which}
\text{the dream-self marries the grand-daughter of John.} \quad \text{(ibid.: 10)}
\]

The ORC now excludes all those structures by means of a concept which P&S
call ‘superiority’ where a starred pronoun \(pro^*\) would have to move across an
unstarred pronoun which a) asymmetrically c-commands it and which b) shares
the same features \(pro^*\) has (cf. (7-d)).
In order to transfer the analysis to counteridenticals and keep the implications it makes with respect to the ORC intact, some adaptations have to be made. First of all, the starred pronoun responsible for dream-self reference is dependent on the left periphery of the embedded CP as the landing site for its lambda abstractor, since it wants to be identified with the center of the speaker’s doxastic worlds (cf. (23)). In the case of counteridenticals, however, no such landing site seems to be available. One way to remedy this problem is to assume an extension of P&S’s theory by Moltmann (2003), who suggests to interpret all propositions as attitudinal objects, e.g.

(24) Mary is happy.
   a. \( \lambda x [R(x,y) \land \langle \text{Happy}, T_1 \rangle, \langle \text{Mary}, T_2 \rangle] \)
   b. An agent predicates, in the assertive mode, the property of being happy \(T_1\) of Mary \(T_2\) (cf. Moltmann 2003: 98)

In this example, \( R \) is an assertion-relation which connects the speaker of the sentence to the propositional constituents, each perceived under a specific mode of presentation, \( T_1 \), of the embedded CP.

For the analysis of counteridenticals, I propose that if sets up an environment in which the proposition \( x \) is \( y \) is taken under a certain attitude; it is interpreted via a ‘counterfactually imagine’-relation, by means of which the speaker predicates, in the cf-imagine-mode, the consequent property to his/her counterfactual-self (cf. (25)). Hence, counteridenticals may be analyzed in the following way: [The speaker (counterfactually) imagines \( [C_P \) that \( (x_i \ be \ y_j, \wedge \ CP(x_i \ y_j))] \),” which clearly incorporates the desired embedded CP.

(25) \[
\begin{align*}
\text{[Preliminary version]} \\
\text{[If } x \text{ were } y \text{]} \ V P(f) \ x w \ w' \ Q(y)(w') = 1. \quad \text{(based on P&S 2003: 8)}
\end{align*}
\]

The predicate \( cf\text{-imagine} \) is analyzed parallel to \( \text{dream} \) in P&S’s account, with the exception that we implement the notion of asymmetric \( \text{be} \) into the meaning of \( CF\text{-IMAGINE}_{x,w} \).

(26) \[
\begin{align*}
\text{[Preliminary version]} \\
\text{CF-IMAGINE}_{x,w} &= \{<y, w'>| \ w' \text{ is a world compatible with the worlds } x \text{ counterfactually imagines in } w, \text{ and } y \text{ is the individual in } w' \text{ from whom } x, \text{ in } w, \text{ takes over a contextually relevant set of properties (meaning that } P(w,y)(w')(x) = 1\} \quad \text{(based on P&S 2003; Percus&Sharvit 2014)}
\end{align*}
\]

Under these assumptions, multiple pronoun reference can be accounted for as in P&S (2003) (cf. (23)), leading to the following analysis of (27).

(27) If I were you, I’d kiss me.
   a. \( I \ \text{cf-imagine} \ [1^* \lambda x \lambda w \ [\forall P \ w_1 \ t_1 \ \text{kiss } \text{me}_2 \ w_1]] \)
   b. \( \forall y \ \forall <y,w'> \text{ in } CF\text{-IMAGINE}_{x,w}, \ y \ \text{kisses } g(2)(w') \text{ in } w' \text{[I]} \)
c. This “property” will hold of the speaker if for all of his/her cf-imagined world, at which s/he takes over contextually relevant properties from the addressee, his/her counterfactually imagined self kisses his/her actual self.

Given the denotation of CF-IMAGINE so far, the proposed analysis can only account for If I were y-conditionals. Nevertheless, a theory of counteridentical meaning should be able to also account for sentences like If Peter were Susan, he would VP and If you were me, you would VP, where the center of the counterfactual worlds is not identified with the person whom the speaker counterfactually imagines to be him-/herself, but rather the person s/he imagines another person to be. For this purpose, I tentatively suggest counteridentical antecedents to attitudinally relate the speaker to a property s/he cf-imagines an entity a to have, whereby a may or may not be the speaker him-/herself (cf. (28)). Since the embedded clause (here: the counteridentical) only attaches below this matrix clause, the lambda abstractor responsible for the interpretation of the starred pronoun, yielding the counterfactual counterpart of a according to the dream-report proposals, can receive its information from a. In other words, I propose that counteridentical antecedents have semantics along the following lines:

\[(28) \quad \text{If } a \text{ were } y\] \[\{x,w,a\} = \{<y,w'>\text{ in } CF-IMAGINE(x,w,a)\}, \text{ whereby} \]
\[CF-IMAGINE(x,w,a) = \{<y,w'>\mid w' \text{ is a world compatible with the worlds } x \text{ counterfactually imagines in } w, \text{ and } y \text{ is the individual in } w' \text{ from whom } a, \text{ in } w, \text{ takes over a contextually relevant set of properties (meaning that } P(w,y)(w')(a) = 1)\}.\]

\[(29) \quad \text{If Susan were Sue, she cf-self'd be in love with her actual self brother.} \]

a. I cf-imagine of Susan [ she* λ3 [ λw1 [ VP w1 t3 be in love with [ her2 w1] brother | | | |] ] ] ] ]  
b. [λa. λx. λw. ∀ <y,w'> in CF-IMAGINE(x,w,a), y kisses g(2)(w') in w'(Susan)(I)  
c. This “property” will hold of Susan if for all of the speaker’s cf-imagined world, at which Susan takes over contextually relevant properties from Sue, her counterfactually imagined self is in love with her actual brother.

It is left to future research to investigate the compatibility of this proposal with existing theories of attitude verbs, as well as the reason why the principle B violations persist in the case that a ≠ y (cf. section 1, similarity 2.). Finding answers to these questions might also help to shed further light on the integration of the analyses of counteridenticals and dream reports, since reports like I dreamed that Peter was Sue and that he married his brother seem parallel to counteridenticals with second and third person subjects, but have not been included in P&S’s (2003) theory.

One tentative argument in favor of an analysis which incorporates two different centered worlds – i.e. a world centered around the subject and one centered
around its non-actual, composed counterpart – is that it predicts the duality of deixis observed in counteridenticals: Whereas some indexicals seem to always be anchored to the speaker (cf. (30)), others seem to be relative to the counterfactual counterpart (cf. (31)). Note that in the examples, the relevant deictic center has been made explicit by means of subscripts. Interestingly, the observed deictic relations persist regardless of the entities identified with each other by means of the antecedent clause.

(30) a. If I were Mary, I wouldn’t be dating that horrid guy\attitude{speaker}\.
    b. If Peter were Mary, he wouldn’t be \underline{here}_{\text{speaker}} \underline{right now}_{\text{speaker}}.

(31) /Assuming Mary is at the beach in Spain./
    a. If I were Mary, I would taste all of the local\local_{\text{Mary}} \underline{goodies}.
    b. If Peter were Mary, he’d jump into the sea \underline{in front of him}_{\text{Mary}}.

3 Conclusion

This paper proposes a novel analysis of counteridenticals which takes into account their grammatical and conceptual parallels with dream reports instead of building on Lewis’ counterpart theory. In contrast to existing proposals, it correctly predicts the constraints counteridenticals impose on their consequent pronouns and it is further able to account for the different compositions of the counterfactual counterpart shared by the antecedent clause’s subject and object.

References

Uniformity Motivated

Cameron Domenico Kirk-Giannini

Rutgers University

Abstract. Can rational communication proceed when interlocutors are uncertain which contents utterances contribute to discourse? An influential negative answer to this question is embodied in the Stalnakerian principle of uniformity, which requires speakers to produce only utterances that express the same content in every possibility treated as live for the purposes of the conversation. The principle of uniformity enjoys considerable intuitive plausibility and, moreover, seems to follow from platitudes about assertion; nevertheless, it has recently proven controversial. In what follows, I defend the principle by developing two arguments for it based on premises reflecting the central aims and assumptions of possibility-carving frameworks for modeling inquiry — that is, frameworks which describe the evolution of individuals’ attitudinal states in terms of set-theoretic operations defined over a domain of objects representing possibilities.

1 Introduction

Consider the following two claims about assertion:\(^1\)

(Assertion Rule) If accepted, an assertoric utterance changes the context by adding its content to the common ground.

(Uniformity) In cases of rational communication, an assertoric utterance expresses the same proposition in each possible world in the context set.

(Assertion Rule) renders formally tractable the platitude that assertion is a species of literal linguistic communication: a speaker asserts a content, and, if all goes according to plan, her audience comes to take that same content for granted for the purposes of the conversation. What is more, it is tempting to think that (Uniformity) follows from (Assertion Rule). For, temptation suggests, if (Uniformity) were violated, interlocutors would not know how to apply (Assertion Rule). Stalnaker (2009) has this sort of argument in mind when he writes:

“To allow assertions with content to which the players of the [language] game do not have access is [to] allow a situation in which the players will be unable to apply the rule. It would be like a card game in which [you are] dealt a card, face down, and the rule requires that you draw

\(^1\) See e.g. Stalnaker 1978, 2009, 2014.
another card if and only if the card you were dealt is red. If you are not allowed to look at the card you are dealt, then you [are] not in a position to apply the rule.” (Stalnaker 2009, p. 407)

If (Uniformity) follows from (Assertion Rule), and if (Assertion Rule) approximates the status of a platitude about assertion, then there is a strong case to be made for (Uniformity). But, tempting though it may be, the inference from (Assertion Rule) to (Uniformity) is controversial. Why think that violations of (Uniformity) must put interlocutors in a position from which they are unable to apply (Assertion Rule)? Hawthorne and Magidor (2009, 2011), for example, have argued that if individuals can sometimes come to be belief-related to a content by accepting a sentence that expresses it, then interlocutors can sometimes update the common ground with the content expressed by an assertion even in cases of (Uniformity) violation. To the extent that this is a reasonable account of what it takes to come to believe a content, there is reason to doubt that cases in which (Uniformity) is violated are ipso facto cases in which interlocutors cannot apply (Assertion Rule), and so reason to doubt (Uniformity) itself.

Hawthorne and Magidor’s exchange with Stalnaker thus raises the question of whether there is a compelling argument for (Uniformity) from within Stalnaker’s framework — that is, whether it can be derived from fundamental commitments of that framework rather than a contentious theory of ability. In this paper, I show that there are at least two ways in which the Stalnakerian can argue convincingly for (Uniformity) without appealing to a controversial theory of ability.

2 From (Assertion Rule) to (Uniformity)

In this section, I offer a formal reconstruction of Stalnaker’s argument from (Assertion Rule) to (Uniformity) and Hawthorne and Magidor’s criticism of it.

Stalnaker’s work on assertion in inquiry stands within an influential tradition in the study of content-bearing mental states, which models the contents of those states as sets of entities (points): each point represents a distinct possibility, and each set of points represents the content that is compatible with all and only those possibilities which correspond to its members.2 Such possibility-carving frameworks are useful in that they permit facts about changes in individuals’ attitudinal states to be described set-theoretically, leading to particularly elegant characterizations of processes like inquiry, hypothetical reasoning, and plan formation.

In the Stalnakerian framework, every conversation is associated with a set of participants or interlocutors. These interlocutors perform assertoric speech acts, which are governed by conversational rules. It will be helpful to separate

---

2 I refer to contents rather than propositions throughout so as not to suggest that I take the possibilities represented by points to be metaphysically possible worlds. In the interest of maximal generality, this is an issue on which I wish to avoid taking a stand.
out two aspects of each such speech act: the *sentence* assertorically uttered and the *content* (set of points) asserted. The same sentence, uttered assertively, can constitute the assertion of different contents in different contexts.

Central to the Stalnakerian framework is the notion of *presupposition*. Presupposition is a theoretically primitive attitude interlocutors bear to contents; an interlocutor presupposes a content just in case she takes it for granted for the purposes of the conversation. The set of contents presupposed by every interlocutor in conversation $c$ at point $w$ is the *common ground* in $c$ at $w$ (henceforth $CG_w^c$); the set-theoretic intersection of these contents is the *context set* of $c$ at $w$ (henceforth $CS_w^c$). $CS_w^c$ thus represents the set of possibilities treated as live for the purposes of the conversation.

I will primarily be concerned in what follows with the interactions between the contents of assertions and the context sets of the conversations in which they are performed. Let us call the set of points formed by intersecting the context set of conversation $c$ at point $w$ with the content of an assertoric utterance $u$ at point $w'$ the *CS-content* of $u$ at $w'$ (when the values of $c$ and $w$ are not important, we can also speak of *CS-content* more generally). Since the context set plays a distinctive role in Stalnaker’s theory of inquiry, failures of (Uniformity) which involve CS-content are especially problematic. Indeed, in what follows I will restrict my attention to a version of the uniformity principle which concerns CS-content exclusively.\(^3\)

Having introduced these Stalnakerian concepts, we can now give a formal characterization of the uniformity principle which makes precise the sense in which it is concerned with the CS-content of an utterance:

(Uniformity) In cases of rational communication: for all conversations $c$, assertoric utterances $u$, contents $p$, and points $w, w', w''$: if $w'$ is in $CS_w^c$ and $p$ is the $CS_w^c$-content of $u$ in $w'$, then for all $w'':$ if $w''$ is in $CS_w^c$, then $p$ is the $CS_w^c$-content of $u$ in $w''$.

As we have seen, Stalnaker’s argument from (Assertion Rule) to (Uniformity) relies on motivating an intuition to the effect that in cases where (Uniformity) is violated, interlocutors are not in a position to add the content of an assertion to the common ground. We can render this claim explicit as follows:

(Stuck) For all conversations $c$, assertoric utterances $u$, contents $p$, and points $w, w', w''$: if $w'$ is in $CS_w^c$ and $p$ is the $CS_w^c$-content of $u$ in $w'$, and there is some $w''$ such that $w''$ is in $CS_w^c$ and the $CS_w^c$-content of $u$ in $w''$ is not $p$, then some interlocutor in $c$ is not in a position at $w$ to apply (Assertion Rule) to $CG_w^c$ in response to $u$.

Stalnaker’s argument also requires a principle linking what counts as rational communication on the part of a speaker with the abilities of her audience. The

---

\(^3\) Hawthorne and Magidor (2009, p. 384) call this version of the principle *(Weak Uniformity)* and note that there is textual evidence that Stalnaker prefers it to the unrestricted version of the principle.
intuitive idea is that a speaker ought not to assert if her audience cannot respond appropriately:

(Fair Play) In cases of rational communication: for all speakers s, conversations c, assertive utterances u, and points w: s performs u at w only if every interlocutor in c is in a position at w to apply (Assertion Rule) to CGw in response to u.

(Stuck) and (Fair Play) together secure the truth of (Uniformity).

Consider now the following scenario, a version of which is discussed by Hawthorne and Magidor and which provides a good illustration of the intuition behind Stalnaker’s argument:

Fire: Smith is watching Jones through a doorway. She can see the corridor in which Jones is standing, but not the room into which Jones is looking, though she presupposes, and Jones presupposes that she pre-supposes (and so forth), that the room contains either Bill or Ben and no one else. Bill or Ben (whomever it is) is performing a dangerous chemical experiment. Something goes horribly awry, and Jones turns to Smith and exclaims ‘He is on fire!’.

Let us refer to the point representing the situation described in Fire as w and the conversation between Smith and Jones as c. Since Smith does not know whether the referent of Jones’s ‘he’ is Bill or Ben, there are distinct points w', w'' in CS_w immediately after Jones’s utterance u such that in w' Jones refers to Bill and in w'' Jones refers to Ben. Suppose that in w', u expresses a content which is true just in case Bill is on fire, and that, in w'', u expresses a content which is true just in case Ben is on fire. Because CS_w contains both points at which Bill is on fire and points at which Ben is on fire, the CS_w-content of u at w' is different from its CS_w-content at w''. By (Stuck), then, some interlocutor (here, Smith) is not in a position at w to apply (Assertion Rule) to CG_w in response to u — which of the two contents of u is she supposed to add to the common ground? It follows, by (Fair Play), that Jones’s utterance u does not count as a case of rational communication.4

Hawthorne and Magidor (2009, 2011) seek to undermine the inference from (Assertion Rule) to (Uniformity) by rejecting (Stuck). In the case of Fire, what this claim amounts to is that Smith is able to apply (Assertion Rule) to the common ground in response to u by accepting the sentence ‘He is on fire’ and thereby coming to believe its content, whatever that happens to be.

3 The First Argument

My arguments for (Uniformity) depend, first of all, on an ancillary premise which I take to be fairly uncontroversial. It ensures that, despite the possibility of violations of (Uniformity), the negation operator behaves as it does in propositional

4 Stalnaker (1978) therefore proposes that speaker and audience perform a process of reinterpretation to recover a uniform content for the utterance.
logic — an assertoric utterance of a sentence always expresses the complement of the content expressed by an assertoric utterance of its negation:

(Negation) For any assertoric utterance $u$ of sentence $\phi$ in conversation $c$ at point $w$, if $u$ expresses contents $p_1, ..., p_n$ at different points in $\text{CS}_C^w$, then there is a sentence $\neg \phi$ such that, for each point $w'$ in $\text{CS}_C^w$, if $u$ expresses $p_n$ at $w'$, then an assertoric utterance $u'$ of $\neg \phi$ at $w'$ expresses $W/p_n$.\(^5\)

My first argument for (Uniformity) also relies on two substantive premises. Both are motivated by the same central assumption of the Stalnakerian framework: that it is acceptable for an interlocutor to assert a content just in case doing so would result in narrowing the context set without eliminating it entirely. Thus Stalnaker:

“One cannot normally assert, command, promise, or even conjecture what is inconsistent with what is presupposed. Neither can one assert, command, promise or conjecture what is itself presupposed. There is no point in expressing a proposition unless it distinguishes among the possible worlds which are considered live options in the context.” (Stalnaker 1970, p. 280).

Let us define a function $F$ from assertoric utterances $u$ to sets of points such that, for each $u$, $F(u)$ is the set of points at which $u$ expresses a truth — that is, the set of points $w$ such that $w$ is an element of the assertoric content of $u$ in $w$.

The two substantive premises of my argument describe the relationship between the assertoric content of $u$ and $F(u)$. The first of these premises formalizes the idea that accepting an assertion does not entitle interlocutors to rule out certain possibilities which, for all they presuppose, are compatible with its content:

(Stability) For any conversation $c$, assertoric utterance $u$, and point $w$, if the intersection of $F(u)$ and $\text{CS}_C^w$ is $\sigma$, then applying (Assertion Rule) to the content of $u$ at $w$ does not result in eliminating any member of $\sigma$ from $\text{CS}_C^w$.

The second premise formalizes the idea that accepting an assertion of a given sentence requires interlocutors to rule out all possibilities at which an assertion of the negation of that sentence would express a truth:

(Contraction) For any conversation $c$, assertoric utterance $u$ of a negated sentence $\neg \phi$, and point $w$, if the intersection of $F(u)$ with $\text{CS}_C^w$ is $\sigma$, then applying (Assertion Rule) to the content of an assertoric utterance of $\phi$ at $w$ results in eliminating every member of $\sigma$ from $\text{CS}_C^w$.

\(^5\) Here $W'$ denotes the set of all points in the model and $W/p_n'$ denotes the complement of content $p_n$ in that set.

\(^6\) I assume bivalence, the principle of excluded middle, and that an utterance expresses a content at every point in the context set.
With the aforementioned Stalnakerian assumption in hand, suppose (Stability) were violated — suppose, that is, that it were possible to perform an assertoric utterance \( u \) and thereby eliminate some point in \( F(u) \) from the context set. Consider a case in which every point in the context set is a member of \( F(u) \) (in Stalnakerian terms, a case in which it is presupposed that \( u \) expresses a truth). If the goal of assertion is to narrow the context set, and if this can be done by uttering \( u \), as is predicted given (Assertion Rule) and our assumption that (Stability) is violated, our verdict must be that \( u \) would be a felicitous contribution to joint inquiry.\(^7\)

Considering a case with the required structure makes it clear that this prediction is false. We can modify Fire to provide a familiar example: suppose Smith presupposes not only that either Ben or Bill is in the room but also that whoever is in the room is on fire (and that Jones presupposes that Smith presupposes this, and so forth). Then it is presupposed that Jones’s utterance of ‘He is on fire’, where ‘he’ is intended to pick out the person in room, will express a truth. If (Stability) could be violated in this case, then Jones’s utterance could nonetheless eliminate some point in the context set and thereby inform Smith about some feature of the world. But this is not the intuitive judgment about the case — without straining, one can only imagine Smith rejecting Jones’s utterance as infelicitous (‘Duh!’/ ‘That’s so helpful.’/ ‘Who, damn it?!’). The most straightforward explanation for this fact is that (Stability) holds.

Similarly, suppose (Contraction) were violated. Then, given the connection between the context set and admissible linguistic behavior, there would be cases in which it was presupposed that an utterance of \( \neg \phi \) would express a truth, and yet asserting \( \phi \) would be felicitous because it would narrow the context set without eliminating it entirely.\(^8\) Let us again modify Fire — this time so that what is presupposed is that whoever is in the room is not on fire. If Jones were to assertorically utter ‘He is on fire’, her utterance would be infelicitous (‘What are you talking about?’). The most straightforward explanation for this fact is that (Contraction) holds.

In addition to (Stability) and (Contraction), my first argument will make use of a lemma concerning the interaction between the function \( F \) and the negation operator. The proposition to be established is that, for any two assertoric utterances \( u \) of sentence \( \phi \) and \( u' \) of its negation \( \neg \phi \), conversation \( c \), and point \( w \), if any point in \( CS^c_w \) is not in \( F(u) \), then it is in \( F(u') \). This can be shown as follows: Suppose our point, \( w' \), is not in \( F(u) \). Then, by the definition of F, \( u \), as performed at \( w' \), does not express a truth. So \( w' \) is not an element of the content \( p \) expressed by \( u \) at \( w' \). By (Negation), \( u' \) expresses \( W/p \) at \( w' \). So \( w' \) is

\(^7\) In principle, one could reject (Stability) while banning such cases by introducing some further principle to the effect that (Stability) must be satisfied whenever every point in the context set is a member of \( F(u) \). But I can see no way of motivating such a principle.

\(^8\) Again, one could in principle render problem cases consistent with the falsity of (Contraction) by stipulation, but I will disregard this possibility.
an element of the content expressed by \( u' \) at \( w' \). So \( u' \) as uttered in \( w' \) expresses a truth at \( w' \). It follows that \( w' \in F(u') \).

(Stability) and (Contraction) enable us to give an argument for the following principle:

**Diagonal** For any conversation \( c \), assertoric utterance \( u \), and point \( w \), the \( \text{CS}_c^w \)-content of \( u \) in \( w \) is the intersection of \( \text{CS}_c^w \) with \( F(u) \).

The argument is as follows:

Suppose we have a violation of (Diagonal): a case in which there is an assertion \( u \) at a point \( w \) in a conversation \( c \) such that the \( \text{CS}_c^w \)-content of \( u \) in \( w \) is not the intersection of \( \text{CS}_c^w \) with \( F(u) \) (call this set of points \( \sigma \)). Then the \( \text{CS}_c^w \)-content of \( u \) in \( w \) is either a proper subset of \( \sigma \), or it contains points not in \( \sigma \). If the former, then, by (Assertion Rule), accepting \( u \) in \( w \) would result in eliminating some points in \( F(u) \) from \( \text{CS}_c^w \) in violation of (Stability). If the latter, then by our lemma, the \( \text{CS}_c^w \)-content of \( u \) in \( w \) contains at least one point in \( F(u') \), where \( u' \) is an assertoric utterance of the sentence \( \neg \phi \), and accepting \( u \) in \( w \) would fail to rule out this point, violating (Contraction).

So there can be no violation of (Diagonal).

**4 The Second Argument**

My second argument for (Uniformity) relies on three premises in addition to (Negation). The first of these guarantees that there are sentences in our language which can be used uniformly to express certain contents:

**Sentential Plenitude** For any conversation \( c \), content \( p \), point \( w \), and assertoric utterance \( u \) at \( w \), for each point \( w' \) in \( \text{CS}_c^w \), if \( u \) expresses \( p \) at \( w' \), then there is a sentence \( \psi \) such that an assertoric utterance \( u' \) of \( \psi \) at \( w \) expresses \( p \) at every point in \( \text{CS}_c^w \).
The idea behind (Sentential Plenitude) is that, if there is an assertoric utterance which violates (Uniformity), there must also be corresponding assertoric utterances which do not violate (Uniformity) and which express the various contents that could, given the context set, be expressed by the original assertion. This will usually be the case when (Uniformity) is (putatively) violated because of the use of a context-sensitive expression, for there will usually be context-insensitive expressions which have each of the possible semantic values of the original.

The second additional premise guarantees that, in cases of (Uniformity) violation, interlocutors do not presuppose that the context set depends on which assertoric content an utterance expresses.

**Independence** For any conversation $c$, point $w$, and assertoric utterance $u$, if $u$ violates (Uniformity) at $w$, then for each content $p$ expressed by $u$ at a point in $\text{CS}^u_w$, there is a point $w'$ in $\text{CS}^u_w$ such that (i) $\text{CS}^u_w = \text{CS}^u_{w'}$ and (ii) $u$ expresses $p$ at $w'$.

(Independence) assures us that, in cases of (Uniformity) violation, it is never presupposed that the context set would have to be different than it in fact is for the speaker to express one of the potential contents of her utterance. This principle holds in intuitive cases of (Uniformity) violation; in *Fire*, for example, it would be odd for either Smith or Jones to presuppose that if it were Bill in the room, the common ground would have to be different than it actually is (likewise for Ben).

Finally, the argument relies on a Stalnakerian principle which explicates the intuitive relationship between what is taken for granted in a conversation and which epistemic possibility claims are felicitous continuations of that conversation:

**Correspondence** For any conversation $c$, sentence $\phi$, and point $w$, if it is acceptable for some interlocutor in $c$ to sincerely utter the sentence \("\phi\) at $w$, then there is a point in $\text{CS}^u_w$ which is an element of the content of an assertoric utterance of $\phi$ at $w$.

(Correspondence) enforces a connection between what counts as acceptable linguistic behavior for the participants in an inquiry and the ways in which that inquiry can be modeled, precluding particular modeling choices which do not respect this connection. For example, as Allan Gibbard (ms) has shown, the kind of permissive view of coming to believe a content which Hawthorne and Magidor suggest leads quickly to the conclusion that inquiry is a trivial enterprise. For it is a feature of the linguistic competence of a normal English speaker that she

---

9 Gibbard’s argument concerns individual propositional attitudes rather than the context set and attempts to demonstrate that, if knowing a proposition under any guise is sufficient for knowing it *simpliciter*, speakers can know any true proposition *a priori*. The similarity between Gibbard’s argument and the argument given here, however, should be apparent.
knows that the sentence ‘The world is like this’ expresses a truth as uttered in any context. If accepting a sentence as uttered on a particular occasion is sufficient for ruling out the points incompatible with its content, then one way for conversational participants to narrow the context set to the unique point representing the way things actually are is to utter this sentence, intending with ‘this’ to rigidly designate all that is, and then accept it; the content expressed by ‘The world is like this’, uttered with the aforementioned intention, excludes every point except the point representing the world of utterance.

Intuitively, no model which allows speakers to coordinate on a single point so easily could provide insight into inquiry as it actually occurs. (Correspondence) solves this sort of problem, since it ensures that if it is acceptable for interlocutors to sincerely utter a variety of ‘might’-claims even after accepting the sent

We are now in a position to construct an argument for (Uniformity) which does not rely on (Stuck). If (Uniformity) is violated, then there will be at least two distinct contents expressed by a given utterance at different points in the context set. For any two such contents, we can distinguish two types of cases: cases in which the two contents are not ordered by strength, and cases in which they are. Before proceeding to the formal proof, it will be helpful to consider an example of each type.

Hawthorne and Magidor’s FIRE provides a convenient illustration of the first kind of case. Recall that Hawthorne and Magidor suggest that by accepting Jones’s assertion, Smith can come to believe whatever content was in fact expressed: if Bill is in the room, Smith comes to believe that Bill is on fire, while if Ben is in the room, she comes to believe that Ben is on fire. In either case, they argue, there is no reason to believe that Smith is not in a position to apply (Assertion Rule) to the common ground of the conversation: (Stuck) is false.

But Hawthorne and Magidor’s account of FIRE does not obey (Correspondence). For suppose it is Ben in the room. Then, according to Hawthorne and Magidor, when Smith accepts Jones’s assertion, she thereby rules out all points in the context set at which Ben is not on fire. But it seems reasonably clear that, even after she accepts Jones’s assertion, it is acceptable for Smith to affirm the sentence ‘It might be the case that Ben is not on fire’. So, by (Correspondence), there is a point in the context set which is an element of the assertoric content of ‘Ben is not on fire’ as uttered by Smith or Jones — that is, a point where Ben is not on fire. It follows that Smith both does and does not rule out all points in the context at which Ben is not on fire: Contradiction.

As an illustration of the second kind of case, consider the following scenario:10

**Birthday**: Arabella and Barbarella are wrapping presents on Thursday night for their daughter Cinderella’s birthday on Friday. They plan to present Cinderella with half of her presents on Friday morning and the other half on Friday evening. Arabella leaves for some time to take a

---

10 Here I assume that quantifier domain restriction is a semantic phenomenon, with the domain of quantification determined by the speaker’s intentions. Those who accept accounts of quantifier domain restriction incompatible with these assumptions will have to devise a different illustration of the second kind of case.
work-related phone call. When she returns, Barbarella asserts, ‘All the
presents are wrapped; we can go to sleep.’ It is common ground between
the two that Barbarella either intends to convey (only) that all the morn-
ing presents are wrapped or to convey that all the morning and evening
presents are wrapped, but it is not common ground between the two
which of these two options obtains.

Let us call the conversation between Arabella and Barbarella \(c\) and the point
which represents the situation described in Birthday \(w\). The two contents Bar-
barella expresses at various points in \(\text{CS}_w^c\) are ordered by strength: the proposition
that all the morning and evening presents are wrapped entails the propo-
sition that all the morning presents are wrapped, but not vice versa. Moreover,
even if Arabella accepts Barbarella’s assertion, it is acceptable for her to affirm
the sentence ‘It might be the case that Ben is not on fire.’ It follows, by (Correspondence),
that there is a point in \(\text{CS}_w^c\) which is an element of the assertoric content
sentence ‘It might be the case that Ben is not on fire’. So, by (Assertion Rule),
Magidor, when Smith accepts Jones’s assertion, she thereby rules out all points
which represent the situation described in \(w\).

Before proceeding to the formal proof, it will be helpful to consider an
example. Assume, for definiteness, that Barbarella states, ‘All the
evening presents are wrapped.’ Since Barbarella has just been wrapped,
and Barbarella knows that most of the presents have been wrapped,
we can reasonably assume that Barbarella intends to convey that the
evening presents have been wrapped. Before Barbarella states her
assertion, Arabella has uttered ‘The world is like this.’ So, using only the information
that her utterance violates (Uniformity), we can conclude (somewhat surprisingly) that
in \(w\) Barbarella does not intend to convey that the evening presents have been wrapped.
But the same reasoning can be applied to any point in \(\text{CS}_w^c\) which has the same context set as \(w\). By
Independence, at least one such point must be one where Barbarella intends
to convey that both the morning and evening presents have been wrapped.

Now for the formal statement of the proof:

Suppose we have a violation of (Uniformity): a case of rational com-
munication in which there is an assertoric utterance \(u\) at a point \(w\) in a
conversation \(c\) and distinct points \(w', w''\) in \(\text{CS}_w^c\) such that the \(\text{CS}_{w'}^c\)-
content of \(u\) at \(w'\) differs from the \(\text{CS}_{w''}^c\)-content of \(u\) at \(w''\). Then there
is some set \(\rho\) of points in the context set at \(w\) such that whether the
points \(\rho\) contains are ruled out of the context set by applying (Assertion Rule)
to \(u\) in \(w\) depends on which content \(u\) expresses. Since this is, by
stipulation, a case of rational communication, there is no problem with
\(u\) being accepted. Suppose it is. Then one of two cases must obtain.

Case 1: Neither of the two contents of \(u\) is a proper subset of the other.

By (Sentential Plenitude), there are sentences \(\phi\) and \(\psi\) which uniformly
express the \(\text{CS}_{w'}^c\)-contents of \(u\) at \(w'\) and \(w''\) respectively. So, by (As-
sertion Rule) and (Negation), accepting \(u\) involves ruling out either all
the points in the \(\text{CS}_{w'}^c\)-content of an utterance of \(\neg\phi\) or ruling out all
the points in the \(\text{CS}_{w''}^c\)-content of an utterance of \(\neg\psi\). But even after
accepting \( u \) in \( w \), it is acceptable for interlocutors to affirm both \( \uparrow \text{It might be the case that } \neg \phi \uparrow \) and \( \uparrow \text{It might be the case that } \neg \psi \uparrow \). By (Correspondence) as applied to the first of these sentences, at least one of the the points in \( \text{CS}_{c_w}^\phi \) is in the assertoric content of \( \uparrow \neg \phi \uparrow \) as uttered at \( w \); similarly for the second. But then neither all the points in the \( \text{CS}_{c_w}^\phi \)-content of an utterance of \( \uparrow \neg \phi \uparrow \) nor all the points in the \( \text{CS}_{c_w}^\psi \)-content of an utterance of \( \uparrow \neg \psi \uparrow \) have been ruled out of the context set at \( w \). Contradiction.

Case 2: One of the two \( \text{CS}_{c_w}^\phi \)-contents of \( u \) at \( w' \) and \( w'' \) is a proper subset of the other. Then \( \rho \) is the set obtained by set-theoretically subtracting the smaller (logically stronger) \( \text{CS}_{c_w}^\phi \)-content from the larger (logically weaker). By (Assertion Rule), either all the points in \( \rho \) are ruled out of the context set when \( u \) is accepted in \( w \), or none are. By (Sentential Plenitude), there is a sentence \( \phi \) that uniformly expresses the stronger of the two \( \text{CS}_{c_w}^\phi \)-contents. By (Negation), there is also a sentence \( \uparrow \neg \phi \uparrow \) which uniformly expresses the complement of the stronger \( \text{CS}_{c_w}^\phi \)-content in the context set. Even after accepting \( u \) in \( w \), it is acceptable for interlocutors to affirm \( \uparrow \text{It might be the case that } \neg \phi \uparrow \). So, by (Correspondence), at least one of the points in \( \text{CS}_{c_w}^\phi \) is in the assertoric content of \( \uparrow \neg \phi \uparrow \) as uttered in \( w \). Since such a point must be a member of \( \rho \) if it was not ruled out by accepting \( u \), and since accepting \( u \) must involve either ruling out every member of \( \rho \) or no member of \( \rho \), it follows that accepting \( u \) in \( w \) does not rule out any member of \( \rho \). But then it must be that the weaker content of \( u \) is asserted at \( w \). Since nothing specific about \( w \) has been assumed, this is also true at every other point \( w''' \) in \( \text{CS}_{c_w} \) such that \( \text{CS}_{w'''} = \text{CS}_{c_w} \). But by (Independence), there is at least one such point where the stronger content of \( u \) is asserted. Contradiction.

So there can be no violation of (Uniformity). \( \square \)

I conclude that there is decisive reason for the Stalnakerian to embrace a notion of the content of assertoric speech acts which satisfies (Uniformity).

References

Gibbard, Allan. (Unpublished Manuscript). “Propositions: Not What We Believe?”


Deriving *Even Though* from *Even* *

Gunnar Lund
Harvard University,
gunnarlund@g.harvard.edu

**Abstract.** This paper explores the semantics of the concessive subordinators *even though* and its relatives. Previous proposals for these subordinators fail to derive the truth conditions compositionally. The proposal presented herein derives the concessive inference of these clauses compositionally, using a standard account of *even*. This further has the effect of relating *even though* to the concessive conditional *even if* as well as a proposal for concessive uses of *still*.

1 Introduction

Concessive clauses are adverbial clauses that express some opposition or general incompatibility between the truth of the matrix clause and the subordinate clause. That is, they indicate that there is something strange, unexpected, or odd in both clauses being true. In English, these clauses are introduced by subordinators like *even though*, *although*, and *though*. These may be illustrated in the following examples:

(1) a. The tree didn’t fall even though it was struck by lightning.
   b. Although I’m no chef, my turkey came out great.
   c. I’m going to go to bed, though I really need to finish this paper.

In general, *even though/although/though* $p$, $q$ entails that both $p$ and $q$ are true. Further, it carries the inference that there is some general conflict between the two propositions. When there is no apparent incompatibility between the two being true at once, the construction is infelicitous:

(2) (Context: Harry only ever goes on walks when it’s not raining.)
    # Harry went on a walk even though it’s not raining out.

Concessive clauses are often connected to causal clauses (e.g. *because*). That is, they are thought to be somehow ‘anti-causal’ or ‘incausal’, and this is the line of thinking that most analyses of concessives follow ([11], a.o.). However, concessive clauses are also linked to concessive conditionals (e.g. *even if* in English). Diachronically, the origin of many concessive clauses is the concessive conditional, and it is an overwhelmingly common trend that concessive conditionals become

* I thank Isabelle Charnavel, Kathryn Davidson, Gennaro Chierchia, and Roger Schwarzschild for their helpful comments and suggestions on this work.
true concessives, especially ([12]). This is true for English, where *though* was previously conditional ([11]). It is ideal, then, that an analysis of concessives be related to an analysis of concessive conditionals. In English, an obvious parallel between the concessive conditional *even if* and concessive *even though* can be found in the scalar particle *even*.

In this paper, I will provide an analysis of the concessive conditional *even though* that runs parallel to an analysis of *even if* provided by [6]. This analysis will turn on the scalar particle *even*, which compositionally provides the concessive inference to both concessive conditionals and concessives. Further, such an analysis suggests parallels between concessive conditionals, concessives, and the concessive *still* particle.

2 Previous Analysis of Concessives

The most robust and widely cited analysis of concessive clauses is provided by [11] (henceforth K&S). Their study attempts to describe concessive clauses in relation to causal clauses (e.g. *because*). They claim that there’s a general intuition that causal and concessive constructions are related (i.e. they are somehow opposites of one another), and they attempt to support this intuition.

Evidence for this intuition comes from the following pieces of data (Capitalization represents focal accent):

(3) a. John is NOT unhappy because he has lost a lot of money.
   b. John is not unhappy because he has lost a lot of MONEY, (but because...)
   c. John is not unhappy even though he has lost a lot of money.

The intended reading of (3a) is somewhat marginal. The intonation on (3a) and (3b) is crucial; *not* must receive nuclear accent and everything after this accent must be deaccented. This is opposed to (3b), where there is a fall-rise accent on *money*. Both (3a) and (3b) should be read as having wide-scope negation. They differ in that (3a) entails that John is not unhappy. The sentence in (3b) entails that John IS unhappy, but not for the reason that he has lost a lot of money. K&S (as well as [18]) claim that there is an equivalence (I take it a truth conditional equivalence) between (3a) and (3c), where negation in (3c) has narrow scope, applying only to the matrix clause.

K&S assert that a theory’s ability to account for the equivalence (or seeming equivalence) between (3a) and (3c) should be a “criterion of adequacy” for assessing the theory. As such, they propose the following analysis for *because* $p$, $q$ in (4) and *even though* $p$, $q$ in (5):

(4) a. because $p$, $q$
   b. **Presuppositions**: $P \rightarrow Q$; $p$
   c. **Assertion**: $p \land q$

(5) a. even though $p$, $q$
b. **Presuppositions:** \( P \rightarrow \neg Q; p \)

c. **Assertion:** \( p \land q \)

The presuppositions \( P \rightarrow Q \) and \( P \rightarrow \neg Q \) in (4b), (5b) are “generalizations” of \( p \) and \( q \). If these were not generalizations, the asserted content would contradict the presuppositions (i.e. if (5b) were about just the particulars \( p \) and \( q \), (5c) together with (5b) entails the contradiction \( q \land \neg q \)). For example, the presupposition in (3c) means something like: “normally, when someone loses a lot of money, they are unhappy.” To derive the supposed equivalence between (3a) and (3c), they calculate the meanings as follows, where (3a) and (3c) are represented schematically as (6a) and (7a) respectively:

\[
\begin{align*}
6) \quad & a. \neg(p, q) \\
& \text{b. **Presuppositions:** } P \rightarrow Q; p \\
& \text{c. **Assertion:** } \neg(p \land q) \quad (= \neg p \lor \neg q) \\
& \text{d. } p \land \neg q \quad (p \text{ is presupposed}) \\
7) \quad & a. \text{even though } p, \neg q \\
& \text{b. **Presuppositions:** } P \rightarrow Q; p \\
& \text{c. **Assertion:** } p \land \neg q \\
& \text{d. } p \land \neg q
\end{align*}
\]

However, their view of the relationship between *because* and *even though* is not without issue. First, the causal notion of *because* is (obviously) crucial to an analysis of *because*, but not for *even though*. It is possible to construct wide scope (as in (3a)) readings of causal clauses that are not equivalent to a narrow scope reading of *even though*. Cases of non-equivalence turn on the fact that concessive clauses don’t necessarily involve causation at all, unlike *because*-clauses. This can be seen in (8) and (9) below:

\[
\begin{align*}
8) \quad & a. \text{The restaurant isn’t here even though my map says it should be.} \\
& b. \# \text{ The restaurant is NOT here because my map says it should be.} \\
9) \quad & \text{Context: Previous to today, whenever John leaves for work and it’s sunny out, his neighbor’s cat greets him at the front door. Today, the cat didn’t greet John and it’s sunny out. John says:} \\
& a. \text{The cat did not greet me even though it’s sunny out.} \\
& b. \# \text{ The cat did NOT greet me because it’s sunny out.}
\end{align*}
\]

The (b) sentences are very odd. Likely, this is because scenarios don’t involve causation, only a kind of correlation. In (8), it is understood that maps don’t cause things to be in the places that they are. However, we do generally expect that when a trusted map indicates that something should be somewhere, it will actually be there. Similarly, in (9), it is clear that the sun being out doesn’t cause the cat to greet John. The (b) examples appear to be an utterance by someone dismissing the idea that the subordinate clause would be causally related to the matrix clause.
According to K&S’s truth conditions, however, there shouldn’t be any difference between the (a) and (b) sentences. In particular, their view seems to get causal constructions wrong; a real notion of causality needs to be baked into the theory. The presupposed conditional connective alone isn’t enough to describe causation (see [18] for further discussion). This fact is not necessarily fatal to their analysis of concessive constructions, though. The data in (9), for instance, doesn’t dispute that there may be a presupposition that, normally, if it’s sunny out, the cat will greet John at the door. In fact, this seems perfectly consistent. What this shows, however, is that causation is not the “opposite” of concession.

A more serious concern for K&S’s theory lies in the close relation between concessives and concessive conditionals (especially those with scalar particles) cross-linguistically. [12] notes two pieces of evidence for this. He cites [10] who writes, “In many, and perhaps all languages, concessive conditionals with focus particles can be used in a factual sense, i.e., in exactly the same way as genuine concessive clauses”. Further, historically, concessive conditionals often develop into concessives ([12]), yet the reverse is not true. That is, the transition from concessive conditionals to concessives constitutes a grammaticalization path. This is even true of English, as pointed out by K&S, where though was once a conditional subordinator. From K&S’s view of concessives, there is no obvious or intuitive connection to concessive conditionals. Given the apparent closeness between the two constructions, an ideal theory of concessives has close parallels to a theory of concessive conditionals.

A third problem concerns the compositional nature of K&S’s proposal. The subordinator even though contains the scalar particle even. Again, concessive conditionals with scalar particles specifically become concessives. As such, it might be expected that the particle is doing some work in concessives, and an ideal theory will have an account of this.

My proposal for even though will address these three points. First, it won’t make reference to causation, and more clearly represents the correlative nature of even though. it will be constructed as an extension of a compositional theory of even if adopted from [6]. In this way, the close connection between concessives and concessive conditionals will be established. Further, it will be compositional in that it will make use of a standard view of even and a new semantics for though. As an upshot, it can further explain why bare though carries with it the same concessive inference as even though.

3 Even If

The concessive conditional even if is puzzling due to what [13] calls the “consequent entailment problem”. That is, in an even if-conditional, the consequent may be entailed, unlike in a regular if-conditional where the truth of the consequent is contingent. This can be seen in the following:

(10) a. Even if that bridge were standing, I wouldn’t cross the river (⇒ I won’t cross the river)
b. If the bridge isn’t standing, I won’t cross the river. (≠ I won’t cross the river)

Previous attempts to account for this by [13] and [2] are problematic as they fail to derive the entailment 1) compositionally and 2) in accordance with more general empirical facts about even. More recently, [6] (henceforth G&L) provide a view of even if where an independently motivated even composes with the conditional sentence and the presuppositions of even result in an entailment of the consequent.

3.1 Background on Focus and Even

The view on focus adopted here and in G&L is developed in [14], [15]. The focus associating particle even quantifies over a set of propositions determined by the ordinary value and focus value of its scope. The ordinary value of an expression, represented by $[[\_]]^O$, is whatever the usual semantic value of that expression is (i.e. $[[\text{Mary}]]^O = \text{Mary}$). The focus value of a sentence, represented by $[[\_]]^F$, is the set of propositions with the focused expression in the sentence being replaced by items of the same type (i.e. $[[\text{John likes [Mary]}]]^F$ is the set of propositions John likes $x$ where $x \in D_e$). Rooth argues even has propositional scope at LF. Even then takes the ordinary value of the proposition (the prejacent) and a contextually-specified subset C of the focus value (the p-set). To ensure that C is a subset of the focus value, Rooth uses the $\sim$-operator, where $\alpha \sim C$ presupposes that C is a subset of $[[\alpha]]^F$, $[[\alpha]]^O \subseteq [[\alpha]]^F$, and $\exists X \in [[\alpha]]^F. X \neq [[\alpha]]^O$. According to [9], even introduces two presuppositions: 1) the scalar presupposition that the prejacent is the least likely to be true among the alternatives in C, and 2) the additive presupposition that one of the alternatives in C is also true. Further, the prejacent is asserted to be true. Thus the formal denotation of even is:

\[
[[\text{even}]](C)(p)(w) \text{ is defined iff } \\
\exists q \in C[q \neq p \wedge q(w) = 1] \wedge \\
\forall q \in C[q \neq p \rightarrow p < \text{likely/expected } q] \wedge \\
\text{If defined, then } [[\text{even}]](C)(p)(w) = p(w)
\]

Additive presupposition
Scalar presupposition
Assertion

3.2 G&L’s Proposal

The notion of focus is a crucial aspect of the theory of even sketched above. An even if conditional like that in (10a) has no clear focused constituent. However, G&L note that the accent in (10a) may fall on either the auxiliary (“were”) or the main verb (“standing”) of the antecedent and the consequent is still entailed. Thus they propose that what is being focused is some null operator, as in VERUM focus in [7]. This is a covert operator appearing on the antecedent clause that G&L call AFF, which is an identity function:

\[
[[\text{AFF}]]^O = \lambda \phi. \phi
\]

When this is focused, it will generate only one other alternative: negation of the lambda term. Thus the focus value of AFF is:
Thus (10a), repeated as (14a), has the LF in (14b). The focus value of the prejacent is given in (14c), where the two alternatives are named (a1) and (a2).

(14) a. Even if the bridge were standing, I wouldn’t cross the river. (= p)  
    b. Even(C) [if [AFF]F the bridge were standing, I wouldn’t cross the river]∼C  
    c. \{ if the bridge were standing, I wouldn’t cross the river; (= a1)  
        if the bridge were not standing, I wouldn’t cross the river (=a2) \}  

With this, we can now calculate the contribution of even, i.e. the assertion and presuppositions.

(15) a. Assertion: if the bridge were standing, I wouldn’t cross the river.  
    b. Additive presupposition: ∃q ∈ \{a1, a2\}[q ≠ p ∧ q(w) = 1]  
      ⇔ a2 = if that bridge were not standing I wouldn’t cross the river = 1  
    c. Scalar presupposition: ∀q ∈ \{a1, a2\}[q ≠ p → p <_{likely/expected} q]  
      ⇔ a1 <_{likely/expected} a2  

The additive presupposition here is the crucial element in entailing the consequent. It ensures that the alternative that is not the prejacent is true. Therefore, we have the following true propositions: “If the bridge were standing, I wouldn’t cross the river” and “if the bridge were not standing I wouldn’t cross the river”. So in any case, whether the bridge is standing or not, I wouldn’t cross the river. Further, the concessive flavor is imparted by the scalar presupposition. Namely, it would be more likely that I wouldn’t cross the bridge if it were not standing.

4 The Proposal for Concessives

Concessives differ from concessive conditionals in that the propositions in both clauses are entailed. Using (16) as an example, notice that the propositions in both the matrix and the subordinate clause must be true for the sentence to be felicitous. Along with this is the implication of incompatibility or unlikeliness between the two clauses. I will call this the “concessive flavor”.

(16) Even though it’s raining, John went out for a walk.  

My proposal for even though-clauses will take G&L’s proposal for even if as a starting point. The essential insight of that proposal is that there is some focused constituent that leads to a p-set composed of simply a proposition and its negation which even then operates over. G&L achieved this with the AFF operator. This same tack could be pursued for even though, but it raises a few problems. Unlike if, the meaning of though on its own is fairly nebulous. What would though without even mean? Instead, I will pursue a different theory where though itself is the focused operator.
This requires a particular semantics for *though*. The essence of *though* is much like AFF, but with differences that give it a bit more elegance and allow for further predictions. The ordinary value of *though* is just the identity function defined for truth values:

\[(\text{though})^O = \lambda \varphi. \varphi\]

The focus value of *though* will then be a set containing the ordinary value and the negation of the lambda term:

\[(\text{though})^F = \{\lambda \varphi. \varphi, \lambda \varphi. \neg \varphi\}\]

So far, this is the same as AFF. However, I propose that *though* obligatorily introduces a set of alternatives, much in the way that [4] proposes for polarity sensitive items like *any*. As such, *though* is always carrying a kind of F-marking; alternatives are always active when *though* is present. These alternatives must be exhaustified, in this case by *even*.

*Though*-clauses also compose with the matrix clause differently. In a way, *even though*-constructions behave somewhat like conjunction, as the propositions in both clauses are entailed. However, the *even though*-clause is, syntactically, an adjunct, not a conjunction. We can derive logical conjunction via Predicate Modification, much like other adjuncts. *Even* will then take scope over the two conjuncts leading to an LF like the following for (16):

\[(19) \quad \text{Even}(C) [\text{[though}_F \text{ it’s raining]}, \text{John went out for a walk} \sim C]\]

This allows us to make the following derivation:

\[(20) \quad \begin{array}{l}
\text{a. Assertion: THROUGH it’s raining, John went for a walk} \\
\quad = p \land q \\
\text{b. Alternatives: } \begin{cases} p \land q (= a1) \\
\neg p \land q (= a2) \end{cases} \\
\text{c. Scalar presupposition: } \forall q \in \{a1, a2\} \{q \neq p \rightarrow p \lessdot \text{likely/expected } q\} \\
\quad \leftrightarrow a1 \lessdot \text{likely/expected } a2 \\
\quad \leftrightarrow (p \land q) \lessdot \text{likely/expected } (\neg p \land q) \\
\end{array}\]

The scalar presupposition captures the concessive flavor of the construction. In words, (20c) says that John going for a walk given that it’s raining is less likely than him going for a walk when it’s not raining. In general, it rightly derives the reading that it’s unexpected that both the subordinate and matrix clauses would be true at once.

The additive presupposition has been omitted here, crucially different from the proposal for *even if*. In fact, the additive presupposition must be omitted here. If the additive presupposition applied in this case, it would result in contradiction. One alternative, (a1) in (20b) entails \(p\) and the other, (a2) entails \(\neg p\). The additive presupposition of *even* itself is not without controversy. [16] argues against an additive presupposition for *even* on the basis of examples like the following:

\[(21) \quad \text{She’s even an ASSOCIATE professor.}\]
The relevant alternatives here would be something like “She’s an assistant professor”, etc. These alternatives are mutually exclusive; one can’t be true without the others being necessarily false. While there are scenarios where she might be an assistant professor at one institution and an associate professor at another, this isn’t the reading of the sentence. This sentence can be true even when she’s an associate professor at one school and not any other kind of professor. The lack of the additive presupposition in even though, then, is in line with other cases where alternatives are incompatible. This is a controversy about even that I don’t intend to solve here, but I will briefly return to it below in Sect. 5.1.

5 Advantages and Consequences of this Proposal

At the outset of this paper, I noted a certain desideratum of an analysis of even though: it compositionally incorporates the scalar particle even in a way closely resembling even if. In Sect. 5.1 below, I describe this connection in further detail. In addition, I explore bare though in Sect. 5.2 and the related particle still in Sect. 5.3.

5.1 The Synchronic and Diachronic Connection between Even Though and Even If

This proposal makes clear that the formal mechanisms involved in the proposal for concessive conditional even if are similar to that of concessive even though. Both include a focused operator that introduces a set of two alternatives exhaustified by even. They differ in two respects. First, they differ in the composition of the prejacent. In G&L’s analysis of even if, the prejacent is simply an if-conditional, perhaps composed in whatever way the reader sees fit; the covert AFF operator activates the proper alternatives without really affecting the meaning of the prejacent. In my proposal for even though, despite having a similar semantics to AFF, though allows the subordinate clause and the main clause to conjoin via Predicate Modification, resulting in the entailment of the propositions in both clauses.

Second, G&L’s analysis of even if makes crucial use of the additive presupposition. Without the truth of the other alternative, only the prejacent would be entailed, and it is simply an if-conditional where the truth of both antecedent and consequent are contingent. Even though statements, by contrast, can’t involve an additive presupposition. If the additive presupposition were there, a contradiction would arise and we would expect ungrammaticality. As explained above, however, the fact that no additive presupposition arises with even though is in line with other examples involving even and incompatable alternatives. [5] offers a solution for this differing behavior. He argues that even involves two operators EVEN and ADD targetting the same focused constituent. The EVEN operator contains only the scalar presupposition and entails its prejacent. The ADD operator also entails the prejacent and contains both a modified additive
presupposition and the scalar presupposition. The modified additive presupposition in ADD requires only that more likely compatible alternatives be true. Thus, when alternatives are incompatible, there is no ADD and no additivity. The difference in additivity between even though and even if can be explained, then. The prejacent of an even though construction entails that the alternative is false. The prejacent of an even if construction, on the other hand, doesn’t entail that the alternative is false. With the picture of additivity provided by [5] the additive behavior falls out from the relationship between the alternative and the prejacent.

Under this view of even though and even if, the crucial difference between the two is the character of their prejacent. The prejacent of an even though construction is a conjunction of two propositions, whereas the prejacent in an even if construction is a conditional. The difference in additivity falls out as a consequence of the modified presupposition above.

There’s a lot to be said about the diachronic picture that will have to remain unsaid here. Crosslinguistically, concessive conditionals with scalar particles develop into concessives ([12], [10]). Further, some languages (e.g. Italian) utilize a concessive conditional construction for both concessive conditionals as well as concessives. Presumably the distinction is distinguished in context. That is, the construction is taken as concessive when it is already known (or perhaps part of the common ground) that the subordinate clause is true. [3] attribute the change from concessive conditional to concessive to “a hearer’s tendency to infer as much as possible from the speaker”. This inference could in principle be drawn from repeated use of the concessive conditional when the truth of the subordinate clause has been established. Learners then begin to interpret all concessive conditionals as having an entailed subordinate proposition, leading to the concessive. This is a very messy and informal sketch, and much more work needs to be done. However, the view of concessives argued for here is compatible with this line of thinking. Though is essentially a bleached if denoting an identity function.

5.2 Bare Though

So far, this paper has dealt with only even though. English has two other concessive subordinators: though and although\textsuperscript{1}. Though in particular introduces a particular difficulty:

(22) a. Even though it’s raining, John went for a walk.
   b. Though it’s raining, John went for a walk.

The sentences in (22) seem to have essentially the same meaning. The propositions in both clauses are entailed and the concessive flavor (i.e. the scalar presupposition of even) is preserved despite a missing even in (22b). The puzzle, then, is explaining this equivalence.

\textsuperscript{1}I won’t discuss although here, but see [17] for differences between it and even though. It is likely that the scalar presupposition has been lexicalized in the construction.
With the proposal sketched above, the solution is fairly simple. Recall that *though* activates a set of alternatives. Having been activated, they must then be exhausted. The overt *even* exhaustifies these in *even though*, as was shown. Following [4], there are two covert exhaustifiers we can invoke for bare *though*, a counterpart to *even* and a counterpart to *only*. If a covert *even* is present, the result will be equivalent to having the overt *even*.

Our other candidate, covert *only*, however, leads to triviality. I assume that the semantics of *only* and its null counterpart *O* are the following:

\[
\begin{align*}
[\text{only}_C]/[O_C] &= \\
&= \lambda p. \lambda w. p(w) \land \forall q \in C[q \neq p \rightarrow \neg q]
\end{align*}
\]

Simply put, *only/O* affirms the prejacent and negates all other alternatives. When *O* is applied to *though* p, q, the result is trivial:

\[
\begin{align*}
(24) & \quad \text{a. LF: } O(C) \left[(\text{though}_F p), q\right] \sim C \\
& \quad \text{b. Assertion/prejacent: } p \land q \\
& \quad \text{c. Alternatives: } \{ p \land q \text{ (=} a1) \} \\
& \quad \{ \neg p \land q \text{ (=} a2) \} \\
& \quad \text{d. Negation of alternative: } \neg(\neg p \land q) \\
& \quad \text{e. De Morgan’s law & double negation: } p \lor \neg q
\end{align*}
\]

The result in (24e) is entailed by the assertion in (24b), meaning that *O* contributes nothing. As argued for in [1], *only* can’t be vacuous. Since *O* is vacuous in this case, it can’t exhaustify the alternatives of *though*. Thus, of the two covert exhaustifiers, only one, the counterpart of *even*, can grammatically exhaustify the alternatives introduced by *though*. This leaves covert *even* as the only grammatical exhaustifier. Therefore, the equivalence in (22) is the result of a covert *even* in (22a).

### 5.3 Concessive Still

The particle *still* has several uses, including a concessive use. This can be seen in this example, from [8]:

\[
(25) \quad \text{John studied all night. He still failed the test. (from [8])}
\]

Our pre-theoretical understanding of this sentence seems to be much like that of an *even though*-clause. The *still* here indicates some incompatibility between the fact that John studied all night and the fact that he failed the test. [8] provides a semantics for this use of *still* incorporating the scalar presupposition of *even*. *Still* takes two arguments. First, it takes a covert *pro* argument coreferenced with some previous proposition in the discourse, and second, it takes the overt proposition. This makes it type \((<\text{st}>,<\text{st},t>)\). Its denotation (simplified slightly for ease of reading) is the following, with *p* as the *pro* argument:

\[\text{[4] argues that there are significant differences between covert and overt *only*, e.g. in the nature of their presuppositions. These differences are insignificant here, and for the sake of simplicity and space, I will ignore them.}\]
Simply put, only the semantics of even result will be equivalent to having the overt a counterpart to Following [4], there are two covert exhaustifiers we can invoke for bare though for ease of reading) is the following, with ⟨⟨proposition. This makes it type with some previous proposition in the discourse, and second, it takes the overt even exhaustified. The overt though activates a set of alternatives. Having been activated, they must then be incorporating the scalar presupposition of even if. This has the effect of looking very similar to the proposals for even if and even though above.

One issue for this analysis, as I see it, is that the covert pro is stipulated to be a focused constituent. First, this alone would be strange, as I know of no other case where a covert pro is focused. Second, if pro is focused, the set of alternatives should include several other propositions that simply are not the pro argument, not just its negation. Further, still can coexist with even though, provided it appears in the matrix clause.

Alternatively, it’s possible that still itself has a semantics similar to though; it is an identity function that introduces a set of alternatives as well as the focused constituent itself. A covert even would then be the exhaustifier, just as with though. There is no time here to work out the details of this proposal, but certainly an approach like the one just sketched appears promising.

6 Conclusion

I have shown that the meaning of even though can be derived compositionally, incorporating an independently motivated meaning for even and a novel semantics for though. This proposal relates to a semantics for even if presented in [6]. As such, it unifies concessives with concessive conditionals, which are diachronically and crosslinguistically closely related. This proposal further resembles a semantics for concessive still, suggesting further unification.

References

Where Force Matters: Embedding Epistemic Modals under Doxastic Attitudes*

Maša Močnik
Massachusetts Institute of Technology
mocnik@mit.edu

Abstract. This paper documents an existential doxastic attitude verb from Slovenian and shows that it does not embed an anchored universal epistemic modal. It discusses the theoretical implications of the data and concludes that there is no satisfying theory yet.

1 Introduction

What does it mean for Sue to believe that it might be raining or for John to know that Bob must be in the office? There has been a recent surge of interest, see Mandelkern (2017, chapter 4) for an overview, in the interaction of epistemic modals and epistemic/doxastic attitudes. Embedded epistemics have been noted to intuitively quantify over the attitude worlds:

(1) Sam thinks it might be raining. (Stephenson, 2007)
≈ it is compatible with Sam’s beliefs that it is raining

(2) Sam thinks it must be raining.
≈ in all of Sam’s beliefs it is raining

A substantial part of the debate, especially in philosophy, has revolved around refining the truth-conditional glosses above. In addition, the recent work of Anand and Hacquard (2013) has sparked an interested in the landscape of epistemic-embedding attitudes, e.g. Crnič (2014) and Ippolito (2017).

I present novel data relating to these two areas. The data comes from the Slovenian verb dopuščati (‘let’), which is cross-linguistically unusual as a plain existential doxastic verb. To illustrate, (3) conveys that John leaves it open as to whether he made a mistake.

(3) Janez dopušča, da se je zmotil.
John allows that refl aux erred
‘John allows for the possibility that he made a mistake.’

* Thanks to Kai von Fintel, Danny Fox, Irene Heim, Roger Schwarzschild, Rafael Abramovitz, Rajesh Bhatt, David Boylan, Cleo Condoravdi, Luka Crnič, Jon Gajewski, Valentine Hacquard, Sabine Iatríðou, Roni Katzir, Justin Khoo, Matt Mandelkern, Viola Schmitt, anonymous reviewers, and the participants of 24.991.
One of the key data points is that dopuščati, unlike think/believe/know, cannot embed an epistemic necessity modal, as illustrated in (4).

(4) Ian and Sue see people jostling for some freshly-baked pie. Ian says:

∗Dopuščam, da mora biti dobro.
I allow that must be good

‘I allow for the possibility that it must be good.’ (intended)

I discuss the details of this claim as well as the general properties of this verb in §2, where I also discuss the findings of Anand and Hacquard (2013). In §3.1 I explain why a purely semantic account cannot work (it does not generate a clash) and in §3.2 I discuss an example of a pragmatic approach, Ippolito (2017), and say what stipulation needs to be added to account for the data in §2. I conclude in §4 with some notes on a different pragmatic route.

2 Embedding Epistemic Modals

2.1 Slovenian Existential Belief

Slovenian has a verb that can be used deontically to mean ‘let’ (as in, He lets us run around) and doxastically to mean ‘allow for the possibility that’ as in (5).\(^1\) The latter conveys that Othello considers it possible that Desdemona loves Cassio, but he leaves it open as to whether or not she actually does.

(5) Othello dopušča, da Desdemona ljubi Cassija.
Othello allows that Desdemona loves Cassio

‘Othello allows for the possibility that Desdemona loves Cassio.’

Some English speakers use allow in a similar way (Othello allows that Desdemona loves Cassio), especially with might in the embedded clause (I allow that I might be wrong).\(^2\) It seems however that allow is much more discursive, used for example to admit something to be true for the sake of the argument. The Slovenian predicate does not carry any such implication – it merely reports whether the agent’s doxastic state leaves something open.

The logically weak contribution of dopuščati can be seen in (6a), which is not contradictory (assuming that one cannot be inside and outside at the same time), unlike (6b) with verjeti (‘believe’).

I allow that is inside and I allow that is outside

‘I allow for the possibility that he’s inside and I allow for the possibility that he’s outside.’

\(^1\) Dopuščati is part of the standard language and is present to varying degrees in the dialects. The judgments reported here are based on the intuitions of speakers from central Slovenia.

\(^2\) Thanks to Matt Mandelkern (p.c.) for first pointing this out to me.
b. "Verjamem, da je notri, in verjamem, da je zunaj.
   I.believe that is inside and I.believe that is outside
   ‘I believe that he’s inside and I believe that he’s outside.’

Furthermore, a dopuščati claim can be strengthened to a belief or a knowledge claim, as in (7). Compare with Some students passed — in fact all of them did.

(7) In a debate with Flat-Earthers, a scientist is asked:
   a. Ali dopuščate, da je Žemlja okrogla?
      Q you.allow that is Earth round
      ‘Do you allow for the possibility that the Earth is round?’
   b. The scientist replies:
      Seveda dopuščam, da je — verjamem, da je!
      of.course I.allow that is I.believe that is
      ‘Of course I allow for it — I believe that it is round!!’

The natural question to ask is whether dopuščati (so far assumed to be an existential quantifier over the doxastic state) is in fact the dual of verjeti. While the lack of contradictoriness in (8) suggests that it might not be, the answer is not crucial for our purposes — what matters is that it is a weakly quantificational predicate of a doxastic kind.²

(8) Marija verjamem, da je notri, ampak dopušča, da je zunaj.
   Mary believes that is inside but allows that is outside
   ‘Mary believes that he’s inside but allows for the possibility that he’s outside.’

On a syntactic note, dopuščati commonly appears with the noun možnost (‘possibility’).³ The latter is however optional and does not contribute a significant meaning difference.

(9) Othello dopušča (možnost), da ga Desdemona varu.
   Othello allows possibility that him Desdemona cheats
   ‘Othello allows for the possibility that Desdemona is cheating on him.’

³ One explanation is that dopuščati quantifies over epistemic worlds, traditionally taken to be supersets of doxastic worlds, in which case no contradiction is expected in (8). Other possibilities include the idea that belief is weak (Hawthorne et al., 2016) or that (8) involves some contextual shift, e.g. I believe p but because I may be wrong I allow that not p. I will leave these options aside and simply assume that dopuščati is a purely doxastic, existential quantifier.

⁴ This distinguishes it from its Russian cognate dopuskat’, which is much more common without a noun. The Russian National Corpus (http://www.ruscorpora.ru/en/index.html, last accessed in May 2017) contains 406 tokens of dopuskatju (‘I allow’) immediately followed by a čto (‘that’) clause, compared to 20 tokens of dopuskaju immediately followed by a noun, of which 1 is ‘possibility’ and 19 are ‘thought’ (Rafael Abramovitz, p.c.).
2.2 Embedding Epistemics in Slovenian

In this section I discuss the core facts of this paper – embedding epistemic modals under dopuščati. The upshot is that embedding a possibility modal is acceptable and feels superfluous, whereas embedding a necessity modal is bad when the modal is anchored to the attitude holder.

The Slovenian modal system is highly sensitive to polarity. The necessity modal verb morati (‘must’) is restricted to upward entailing environments, which is fortunate since it can appear in the complement of dopuščati. This is not so for moči (‘may’/‘might’), so instead I will use the adverb mogoče (‘maybe’), which is typically used for the English might.\(^5\)

Embedding an epistemic possibility expression like mogoče (‘maybe’) under dopuščati does not seem to create much of a meaning difference. To illustrate, the following two feel roughly equivalent:

(10) Context: Othello is asked whether he thinks that D is cheating on him.

a. Dopuščam, da me vara.
   ‘I allow for the possibility that she is cheating on me.’

b. Dopuščam, da mogoče vara.
   ‘I allow for the possibility that she might be cheating on me.’

It is questionable whether there is any tangible difference between (10a) and (10b). Speakers do report examples like (10a) to be slightly weaker in comparison, but this difference might be due to the fact that there is more overt material in (10b), compare John might come with John might possibly come. The intuition remains regardless of whether the possibility noun appears overtly.

In contrast, embedding a necessity modal under dopuščati feels strange in the same context (assuming that nobody previously utters the modal claim in the embedded clause).\(^6\)

(11) Context: as (10).

* Dopuščam, da me mora varati.
   ‘I allow for the possibility that she must be cheating on me.’

The oddity remains in a context like (4), repeated below, where a must-claim is otherwise appropriate, as in (12b).

(12) Ian and Sue see people jostling for some freshly-baked pie. Ian says:

---

\(^5\) I stick to mogoče (‘maybe’) since it is epistemically very neutral (as opposed to the verb utegniti, which might better be modelled with a richer semantics involving degrees of likelihood/possibility).

\(^6\) The intuition remains if the verb is in third person with Othello as the subject.
I allow that must be good
‘I allow for the possibility that it must be good.’ (intended)

b. Dobro mora biti.
good must be
‘It must be good.’

c. Verjamem, da mora biti dobro.
I believe that must be good
‘I believe that it must be good.’

Unlike (12c), the sentence in (12a) sounds nonsensical because the agent is intuitively felt to be leaving it open as to whether the pie is good while at the same time being convinced, based on his evidence, that it is good. Recall that no such clash seems to occur in (7).

This clash between an embedded necessity modal and dopuščati disappears when the embedded modal has a different flavour, for example a deontic one:

(13) I was ordered to do something but I can’t remember what.

Dopuščam, da moram iti v trgovino.
I allow that I must go to shop
‘I allow for the possibility that I must go to the shop.’

My intuition is that acceptability also improves when the embedded necessity modal is anchored not to the attitude holder but to some other body of evidence, but I have unfortunately not been able to prove this.7

7 The following example is an attempt to anchor the embedded modal to the evidence in the filing cabinet (after Kratzer’s (2009) filing cabinet example in Anand and Hacquard (2013, p. 23)):

(1) John and Sue are the two prime suspects in a murder investigation. They learn that a key piece of evidence has been uncovered and decide to break into the filing cabinet (where the detectives keep all their evidence) to see who needs to flee as neither of them remembers the night of the murder. They know that the collective evidence now strongly (but not decisively) points either to John or to Sue. Before opening the cabinet, John thinks about the detectives and their evidence in the box and tells Sue:

Dopuščam, da moram biti jaz njihov krivec.
I allow that must be I their culprit
‘I allow for the possibility that I must be their culprit.’

The speakers I consulted still found this sentence bad, presumably because they were not able to anchor the embedded modal to the evidence in the filing cabinet (but instead anchored it to the attitude holder, which is expected to yield deviance).
I conclude this section with a side note on the comparison between embedding an existential epistemic expression under dopuščati and embedding it under verjeti (‘believe’). Consider first the contrast in English:

(14) a. Othello believes that she might be cheating.
    b. Othello allows that she might be cheating.

The English speakers who accept (14b) report it to be weaker than (14a). The intuition is that Othello is felt to have more reason to believe that Desdemona is unfaithful in (14a). The same contrast is felt in Slovenian:

(15) a. Othello verjame, da ga mogoče varo.
    Othello believes that him maybe cheats.on
    ‘Othello believes that she might be cheating on him.’
    b. Othello dopušča, da ga mogoče varo.
    Othello allows that him maybe cheats.on
    ‘Othello allows that she might be cheating on him.’

The speakers report Othello to have perhaps some reason for suspecting Desdemona of cheating in (15a), whereas in (15b) Othello might not think that there is any actual reason to suspect her of being unfaithful – it is simply in principle possible that she is. It is difficult to show whether this difference is in fact truth-conditional, so I will leave it aside here.8

2.3 Main Clause Negation

As illustrated in (1) and (2), universal doxastic predicates like think or believe appear to embed existential as well as universal epistemic modals. Anand and Hacquard (2013, fn. 27), whose paper is based on survey data from Romance, find that main clause negation degrades an embedded necessity modal.9 Consider this example for English and Slovenian:10

(16) Context: You, me, and John see Bob go home from work early. We sit down on some couches in front of Bob’s office. John has his back turned to Bob’s door. He puts on some headphones and starts cheating on the latest homework. After a while, Bob, who has a secret entry to his office, which he used to come back, creeps out of his office and comes up behind John’s back. John, still immersed in cheating, does not notice this. I nudge you and whisper, with both of us staring at Bob:

    a. John (clearly) does not think that Bob might be behind his back.

8 Theories such as Yalcin (2007), unlike e.g. Yalcin (2012), do not predict there to be a truth-conditional difference. See Mandelkern (2017) for references on this debate.
9 See also Crnič (2014) and Ippolito (2017, 14, fn. 9) for similar judgments.
10 The example is crucially constructed in such a way that the embedded modal is ‘anchored’ to the attitude holder, i.e. might and must are interpreted with respect to the attitude holder and not the speaker/hearer (because for them Bob is there).
b. ??John (clearly) does not think that Bob must be behind his back.
c. Janez ne misli, da je Bob mogoče za njegovim hrbotom.
   John not thinks that is Bob maybe behind his back
   ‘John does not think that Bob might be behind his back.’
d. ??Janez ne misli, da mora biti Bob za njegovim hrbotom.
   John not thinks that must be Bob behind his back
   ‘John does not think that Bob must be behind his back.’

In contrast to (2), embedding must under think is somehow obstructed by the presence of negation in (16b). While think is a neg-raiser in English, it should be noted that the Slovenian misli (‘think’) is not. The contrast between the English (16a) and (16b) is maintained in the Slovenian (16c) and (16d).

In contrast, negation does not influence the judgments for dopuščati and a similar contrast to above appears:

(17) Context: as in (16).
      John not allows that is Bob maybe behind his back
      ‘John does not allow that Bob might be behind his back.’
   b. ??Janez ne dopušča, da mora biti Bob za njegovim hrbotom.
      John not allows that must be Bob behind his back
      ‘John does not allow that Bob must be behind his back.’

2.4 Other Existential Attitudes

On the basis of survey data from Romance, Anand and Hacquard (2013) argue that emotive doxastics (fear, hope) and dubitatives (doubt), which they analyse as existential doxastic attitude verbs, embed only existential epistemic modals (similarly in Črnč (2014) and Ippolito (2017)).

(18) a. Jean craint que Marie puisse avoir connu son tueur.
    Jean fears that Marie can-SUBJ have known her killer
    ‘John fears that Mary may have known her killer.’
   b. [*]Jean craint que Marie doive avoir connu son tueur.
    Jean fears that Marie must-SUBJ have known her killer
    ‘John fears that Mary must have known her killer.’
    (Anand and Hacquard, 2013, p. 8:45, added *)

While these embedding facts could be due to some special property of these verbs (e.g. defined only if the doxastic state does not settle the prejacent, as in Anand and Hacquard (2013)) the Slovenian data has shown us that a more general explanation is preferable.

11 See Anand and Hacquard (2013, 8:45–8:46) for means and details.
3 Potential Analyses

For reasons of space I limit myself to making two points. I first illustrate on Yalcin (2007) how a purely semantic analysis does not generate the necessary clash for the dopuščati data, and mention the drawbacks of supplementing it with an indirectness presupposition account in the style of von Fintel and Gillies (2010). I then discuss an example of a pragmatic solution – the recent presuppositional account of Ippolito (2017) and say what stipulation needs to be added to account for the data.\textsuperscript{12}

3.1 A Semantic Approach

Yalcin’s (2007) work on epistemic contradictions leads him to propose that attitudes shift the modal base of the embedded modal to their own domain (the doxastic worlds, for example). A similar move is proposed in event semantics by Hacquard (2006), but I will focus on Yalcin’s proposal here for reasons of space. The resulting semantics privileges the quantification of the embedded modal – the embedded modal re-quantifies over the attitude state, which makes the quantification of the attitude vacuous.\textsuperscript{13,14}

More concretely, Yalcin adds an information state parameter (a set of worlds) to the index of evaluation. Epistemic modals are defined as quantifiers over this parameter, whereas attitudes shift it. For example:

\begin{equation}
\text{a. } [\text{must } \phi]^{c,s,w} = 1 \text{ iff } \forall w^\prime \in s : [\phi]^{c,s,w^\prime} = 1 \\
\text{b. } [x \text{ believes } \phi]^{c,s,w} = 1 \text{ iff } \forall w^\prime \in B^w_x : [\phi]^{c,B^w_x,w^\prime} = 1, \text{ where } B^w_x = \text{def} \text{ the set of worlds not excluded by what } x \text{ believes in } w
\end{equation}

In (19a) we see that modals quantify over a given information state s and shift the world of evaluation (from w to w’ above). In addition to the this, attitude predicates, such as believe in (19b), replace a given information state with their own set of worlds (from s to B^w_x above) with respect to which the embedded clause (\phi) is evaluated.

Let us apply this semantics to (4), where Ian says that he allows for the possibility that the pie must be good. Let B^w_i be Ian’s doxastic state at the world of evaluation w, and let \phi abbreviate the embedded non-modalized proposition (‘that the pie is good’). The truth-conditions we obtain for (4) are as follows:

\textsuperscript{12} Ippolito (2017) notes that her account bears some similarity to Crnič (2014). For reasons of space I do not discuss Crnič (2014), whose paper is a commentary on Anand and Hacquard (2013) and leaves it open how to account for sentences like (16b). In turn, Anand and Hacquard (2013) face the problems discussed by Crnič and cannot account for the dopuščati data either.

\textsuperscript{13} Modulo empty belief states, which linguists are happy to rule out independently. See Mandelkern (2017) for discussion though.

\textsuperscript{14} There have been arguments, see Mandelkern (2017), against assuming that the embedded modal quantifies over the whole attitude domain. Mandelkern (2017) proposes a subset constraint while Hacquard (2010, p. 107) assumes (to the same effect) that epistemic modals contribute an ordering source.
\[(\exists w' \in B^w_i : \forall w'' \in B^w_i : [\phi]^{c,B^w_i,w''} = 1}\]

The existential quantification is provided by *dopuščati* and the universal one by the embedded modal (whose domain of quantification is shifted to $B^w_i$ by the attitude). While there is some vacuity in this LF (see fn. 13 and fn. 14), there is nothing inherently contradictory about it (it is in this sense that a purely semantic answer is not sufficient). Yalcin (2007) predicts that the pie is good in all of Ian’s belief worlds. Recall that a similar situation did not lead to an odd claim in (7), where the Earth being round was (in fact) true throughout the scientist’s belief worlds.

One might supplement this account with the idea that a necessity epistemic modal is not embeddable under *dopuščati* because of its evidential nature in the sense of von Fintel and Gillies (2010), who propose that epistemic modals presuppose that the agent’s evidence is indirect (and perhaps it is odd to be doxastically unsure about your evidence in the *must* cases). While this seems like an attractive line of pursuit, Ippolito (2017) and Mandelkern (2017, ch. 3) have recently cast doubt on the presuppositional status of this inference. Since von Fintel and Gillies (2010) persuasively argue against other options (e.g. it being part of the main content), it is unclear what proposal one should assume to be able to work towards a clash.

### 3.2 A Presuppositional Approach: Ippolito (2017)

Ippolito’s account is aimed at the embedding distinctions between verbs like *believe*, on the one hand, and verbs like *fear*, on the other, cf. §2.4. Her account is particularly relevant because she analyses *fear* as a plain existential quantifier over the doxastic state. She builds on Stephenson’s (2007) semantics for modals and attitudes, but adds that *must* and *might* carry presuppositions about the relationship between the agent’s beliefs and the modal proposition.

\[(\exists w' \in B^w_i : \forall w'' \in B^w_i : [\phi]^{c,B^w_i,w''} = 1}\]

Ippolito’s modals are evaluated with respect to a judge parameter $j$ (this will be the attitude holder in our examples) and carry presuppositions about the judge’s beliefs about the modal claim. For example, *It might be raining* presupposes that it is consistent with the judge’s beliefs that it might, according to the judge, be raining.16,17 Modals have the usual truth-conditions: *must*, in brief, requires that the prejacent be true in all the epistemically accessible worlds.

---

15 Thanks to Viola Schmitt (p.c.) for raising this point.

16 It asserts that the judge believes that it might be raining, due to a covert operator.

17 The paper does not explicitly discuss the introspection principles and presupposition projection out of attitude complements. Sentences like ‘x believes/fears $\varphi$’ where $\varphi$
These truth-conditions are coupled with the so-called **Absorption Principle**, which requires the presuppositions in (35) and (36) to not asymmetrically entail (i.e. be strictly stronger than) the assertion that they appear in. Consider the following example, adapted from Ippolito:

(21) *John believes that it might be raining.*

a. presupposes: \( \exists w' \in B^w_j : \exists w'' \in ST^w_j (EP_j(w')) : p(w'') = 1 \)

b. asserts: \( \forall w' \in B^w_j : \exists w'' \in ST^w_j (EP_j(w')) : p(w'') = 1 \)

The sentence presupposes that it is consistent with John’s beliefs that according to John it might be raining. It asserts that John believes that according to John it might be raining. The Absorption Principle is satisfied because (21a) does not entail (21b). In shorthand notation, we can say that (21) presupposes \( D \Diamond p \) and asserts \( B \Diamond p \), where \( D \) is the dual of \( B \).

Ippolito does not analyse *dopuščati* but emotive doxastics like *fear*, which she treats as plain existential quantifiers over the doxastic state (leaving the emotional component aside). We can therefore extend the analysis straightforwardly to Slovenian. The prediction about embedding a necessity modal under *dopuščati* is the same as for *fear*. It is predicted to be bad because the presupposition \( B \Box p \), due to the embedded necessity modal, is strictly stronger than the assertion \( D \Box p \). This is not so in the case of an embedded possibility modal, where the presupposition and the assertion are actually the same (\( D \Diamond p \)). While the presupposition entails the assertion, the assertion also entails the presupposition, hence the presupposition is not strictly stronger than the assertion.

While the approach predicts the basic embedding pattern under *dopuščati* (negation to be discussed), there are a few concerns. Ippolito (2017, p. 11) is aware of the issue that presuppositions and assertions are sometimes the same, stating that a sentence like *The king of France exists* presumes and asserts the same thing as well. Notice, however, that *Does the king of France exist?* is odd in a way that *Does John believe that it must be raining?* is not. Furthermore, there does not seem to be good evidence for the assumed presuppositions. For presupposes \( \psi \), as a whole presuppose \( B \psi \) (cf. Heim (1992)). (Dopuščati is not an exception.) If \( \psi \) is of the form \( \Box \psi \), where on Ippolito’s account \( \Box \psi \) carries the presupposition \( B \Box \psi \), then the sentence as a whole presupposes \( B B \Box \psi \). It seems to me that the approach assumes that \( B B \Box \psi \) reduces to \( B \Box \psi \) and no further. There is a theoretical desire to reduce the analogous truth-conditional content further to account for Yalcin’s (2007) epistemic contradiction data – Stephenson (2007, 590–591) herself uses such a reduction. However, this is in tension with the proposed account where presuppositions and assertions compete in terms of strength. If we always reduced to the innermost modality (i.e. the presuppositional and the truth-conditional content), all presuppositions and assertions would be the same in the positive sentences – having existential force if they contain *might* vs. having universal force if they contain *must*. In particular, embedding a universal modal under an existential attitude would be predicted to be felicitous. An explicit formulation of the account might resolve this tension.
example, it is odd to respond to John thinks it might be raining with Hey, wait a minute, I didn’t know that it was consistent with his beliefs that it might be raining (or the equivalent with dopuščati in Slovenian).

Lastly, there is an additional stipulation that is brought out by the Slovenian data.\textsuperscript{18} While Ippolito mentions that local accommodation can take place when the presupposition and the assertion are inconsistent, we actually need to assume that possibility and necessity modals differ in their accommodation properties. Consider again the negation data in (17). In order to predict that a negated dopuščati embeds an existential modal, we need a way to make the assertion $\neg D\otimes p$ and the presupposition $D\otimes p$ consistent. We can predict the felicity of such a sentence by appealing to local accommodation (under negation). On other other hand, to predict the infelicity of an embedded universal modal under a negated dopuščati, we must not save the inconsistency between the assertion $\neg D\Box p$ and the presupposition $B\Box p$. So the presupposition of the necessity modal must not locally accommodate under negation, but project and yield an inconsistency. I do not know of any cases where two such similar presuppositions would accommodate differently in the same environment (i.e. in the scope of dopuščati).

4 Conclusion and Outlook

In this work in progress I presented novel data on an existential, belief-like predicate from Slovenian and claimed that it does not embed an anchored epistemic necessity modal, supplementing the literature on embedding epistemic modals under attitudes. I used Yalcin (2007) to explain where a purely semantic account runs into trouble and discussed the approach that seems to best capture the data – Ippolito’s (2017) presuppositional approach. I showed that in order to account for all the data one must stipulate different accommodation properties of epistemic necessity vs. possibility modals.

The pragmatic approaches I mentioned and discussed all involve presuppositions (which for some reason or other do not seem to be empirically quite supported). There is another possible line of pragmatic pursuit that I hope to explore more in the future, based on grammatical implicatures and exhaustification in the sense of Fox (2007) and Magri (2009). In a nutshell, the Slovenian equivalent of Jan allows that the pie must be good seems odd because the speaker does not say the logically stronger but contextually (given the introspection principles) equivalent alternative Jan thinks that the pie must be good. The challenge lies in deriving the empirical distinction between possibility and necessity modals in a principled way, as well as extending this competition to other attitudes, like the emotive doxastics.

\textsuperscript{18} The point can be partially made in English with the issues of must under don’t think.
Bibliography


Are Expressives Presuppositional?
The Case of Slurs

Tristan Thommen
Institut Jean Nicod (CNRS, ENS, EHESS)

Abstract. The present paper explores and objects to a reduction of expressivity to presuppositional content, in particular Schlenker’s indexical and attitudinal attempt at doing so (Schlenker 2007). The main claim of the paper is that the projective content associated with slurs that are embedded under filters projects more broadly than the projective content of (at least some) presuppositions under filters. This claim is made in two steps. First, I show that providing such evidence requires controlling for confounds, namely ignorance implicatures and intensionality. Second, I provide pairs of examples (namely (16)-(17)-(18)/(19) and (25)-(26)/(27)) showing that, once these confounds are controlled for, the projective content associated with slurs embedded under filters projects more robustly than the projective content of (at least some) presuppositions. The discussion is focused on the expressive content of the particular case of slurs, but a generalization to other expressives (e.g. "damn", "bastard", "fucking" etc.) or to projective content in general will thus become worth considering.

1 Introduction

According to Potts and others, expressives in general and slurs in particular display peculiar properties that require the introduction of a novel dimension of meaning, independent of other kinds of content (Potts 2007, McCready 2010, Gutzmann 2015). On the other hand, other authors suggested more parsimoniously that provided the acknowledgment of some extra features, the behavior of slurs could be handled in a standard presuppositional framework, with no need for postulating an additional dimension in one’s theories of meaning on top of whatever is independently needed to take care of presuppositions (Maciá 2006, Sauerland 2007, Schlenker 2003, 2007, 2016, Cepollaro and Stojanovic 2016). In particular, Schlenker (2007), responding to Potts (2007), proposes that expressives carry a presupposition which is indexical (evaluated w. r. t. a context), and attitudinal (predicates something of the agent’s mental state).

1 This work owes a lot to many exchanges with Benjamin Spector, to whom I express my gratitude. I also thank François Recanati, Philippe Schlenker, Yael Sharvit, Judith Tonhauser, and three anonymous reviewers, whose diverse remarks and comments greatly benefited to this work.

2 Incidentally, anticipating the potential existence of various perspectival readings, Schlenker also defines such expressive presuppositions as being sometimes shiftable.
The question of whether slurs can be analyzed as presuppositions has received both positive (Maciá 2006, Sauerland 2007, Schlenker 2003, 2007, 2016, Cepollaro & Stojanovic 2016) and negative (Potts 2007, Richard 2008, Davis and McCready 2016) answers in the literature. The data I discuss below supports the negative side, which could constitute positive evidence in favor of use-conditional accounts of slurs and expressivity à la Potts for instance, or maybe in favor of other types of accounts such as Nunberg’s Gricean view (Nunberg 2016), or Hom and May’s radical truth-conditional theory (Hom and May, 2013, 2015).

The main claim of the paper is that the projective content associated with slurs that are embedded under filters projects more broadly than the projective content of (at least some) presuppositions under filters. This claim is made in two steps. First, I show that providing such evidence requires controlling for confounds, namely ignorance implicatures and intensionality. Second, I provide pairs of examples (namely (16)-(17)-(18)/(19) and (25)-(26)/(27)) showing that, once these confounds are controlled for, the projective content associated with slurs embedded under filters projects more robustly than the projective content of (at least some) presuppositions. The discussion is focused on the expressive content of slurs, but a generalization to other expressives (e.g. damn, bastard, fucking etc.) will thus become worth considering, and might shed light on the nature of projective content in general.

In what follows, I focus on Schlenker 2007’s version of a presuppositional account of slurs, because it is to my best knowledge the presuppositional account that is the most likely to derive as wide a projection profile as is needed, in virtue of two additional features: indexicality and attitudinality. Schlenker provides the following lexical entry for the slur “honky” (see Kaplan 2001), with respect to a context (c) and a world (w):

\[(1) \ [\text{honky}] (c)(w) \neq \# \text{ iff the agent of } c \text{ believes in the world of } c \text{ that white people are despicable. If } \neq \#, [\text{honky}] (c)(w) = [\text{white}] (c)(w)\]

Under this analysis, “honky” and “white” or “white person” make the same truth-conditional contributions to utterances in which they appear, and differ with regard to their presuppositional import. Where “white” does not trigger any presupposition, (or at least no presupposition that is relevant to the present discussion) “honky” triggers a presupposition of a particular sort. The presupposition is about the agent of the context (it is indexical), and more specifically, it is about the agent’s attitudes (it is attitudinal).

According to Schlenker, these linguistic properties are sufficient to derive the effects of slurs. The indexical character of the presupposition, together with the assumption that there are shiftable indexicals (Schlenker 2003, Sauerland 2007) would yield the dependency to a particular perspective that slurs and other expressives were noted to display (Potts 2007). And under such a presuppositional view, expressives like “honky” are predicted to follow the same patterns of projection as what is expected from a presupposition that is indexical and

(i.e. the context of evaluation need not be the context of the actual utterance). That property will not be relevant in what follows.
attitudinal. I now turn to an attempt at falsifying that prediction. As the realm of presuppositions is quite diverse and heterogeneous when it comes to projection, I will systematically compare the behavior of slurs to that of both soft (e. g. the factive “know”, or the existence presupposition triggered by “the”) and hard triggers (e. g. “too”), whose import is very difficult, not to say impossible, to accommodate.

2 Projection profile under filters

It appears that the expressive content of slurs projects to the matrix position even in environments where standard presuppositions tend to get filtered. Although presupposition failures usually arise even when the trigger of a presupposition that is not satisfied in the context of utterance is embedded, there are linguistic environments in which presuppositional material interacts with the standard descriptive dimension, to the effect that projection is blocked (the so-called presupposition filters, Karttunen 1973). I start the comparison between slurs and three standard presupposition triggers (“the” “know”, and “too”) under two such filtering environments (disjunctive filters and conditional filters). Facing the need to control for confounds, such as ignorance implicatures and intensionality, I will then consider two more adapted filtering environments.

2.1 Disjunctive filters

Consider the following contrast:

\[ \text{Context: (Take this context to be the default context in which to evaluate all data in the remaining of the paper): none of the participants to the conversation are prejudiced against German people in any way whatsoever, France is not a monarchy, and there is no salient antecedent for “too”.} \]

(2) France is not a monarchy, or the monarch of France is bald.
(3) ?I am not Germanophobic, or my colleagues know that I am. \text{(adapted from Schlenker 2016, p. 47)}
(4) ?I am not Germanophobic, or my colleagues are Germanophobic too.
(5) !I am not Germanophobic, or John is a boche\textsuperscript{3}.

As we see in (2), the presupposition that there exists a (unique) monarch in France can be locally accommodated - that is, roughly its content does not

\textsuperscript{3} Schlenker constructs and discusses a similar example to address precisely that difficulty. In order to stay neutral with regard to the discussed reduction of expressivity to presuppositional content, I use the symbol “!” to mark the expressivity of utterances, and in particular, the presence of an inference about the speakers emotional state that conflicts with the specified context. I will use the symbol “#” for plain vanilla presupposition failures.
systematically project to the matrix and can stay stuck in the embedded phrase, here the second disjunct (see Karttunen 1974, Heim 1983) -, when it is negated in the first disjunct: the presupposition is not inherited by the entire utterance. But in (5), the presupposition that the speaker is Germanophobic seems to be inherited by the whole utterance (or at least, the sentence is very odd), even though it is negated in the first disjunct. (3) and (4) are somewhat less clear, we will understand why in what follows.

Considering only (2) and (5), we observe a first contrast between slurs’ expressivity and presuppositions in a disjunctive filter, contrary to what a presuppositional view of slurs would expect. Can a presuppositional analysis explain this contrast away?

Note that some previous work has noted that expressives can in fact be evaluated just in their embedded position (e. g. Kratzer, 1999; Schlenker, 2003; Potts, 2007). Consider for example (6):

(6) Every member of my family is so racist. I hate it that they won’t accept that I married a white person. It’s so embarrassing that everyone in my family thinks I married a honky.

Naturally, it may be argued, we understand the speaker of (6) as not sharing her family’s racism at all. However, such considerations do not speak against the general observation that there is a clear preference for the projective reading. First, embeddings under attitude verbs are trickier to interpret than it might first appear, because of the potential intervention of perspectival operations that are to the best of my knowledge not quite well understood yet. Second, even if expressives are in some cases evaluated only in their embedded position, there is still an overwhelming preference for the matrix position. That is in itself puzzling, even granting that there is some indexical component in the presuppositional content of slurs.

But as Schlenker notices, his analysis has in fact the resources to explain the contrast observed under disjunctive filters. In virtue of the indexical nature of the presupposition that Schlenker posits for expressives, an oddity might arise in (5), as well as in (3)-(4), and not (2) through the (possibly obligatory) parallel computation of ignorance implicatures: whereas it is conceivable that the speaker does not know whether or not France is a monarchy, it is hard to buy that she does not have access to her own attitudes. As a result, ignorance implicatures are non-problematically derived in disjunctive statements like (2), but they clash

———

4 Potts also makes the observation that expressives display different projection behaviors under the scope of propositional attitude predicates, whereas presuppositions are typically cancelled in these environments (Potts 2007). But such apparent contrasts are not telling because it is in fact not the case that presuppositions are typically cancelled under attitude verbs. The consensus seems to be that presuppositions triggered under attitudes are evaluated in both the embedded position and in matrix position (e. g. Heim, 1992; Zeevat, 1992; Geurts, 1999; Singh, 2008; Beaver and Geurts, 2011; Schlenker, 2011).

5 I thank an anonymous referee for bringing this case to my attention.
with common world-knowledge in disjunctive statements like (5) (see e.g. Magri 2009 for more on that point). As disjunctive statements, utterances of (2)-(5) will undergo the following neo-Gricean enrichment, following Sauerlands (2004) proposal for the computation of scalar implicatures:

(7) Take A and B, two propositions, and K, an unspecified epistemic operator:

- We assume that <A∧B, A, B, A∨B> form a scale
- Utterance: A∨B
- Application of the maxim of Quality: K(A∨B)
- Generation of alternatives: A, B, A∧B

- Primary implicatures: 
  - ¬(K(A)); ¬(K(B)); ¬(K(A∧B))
  - Secondary implicatures: 
    - ¬(K(¬(A))); ¬(K(¬(B)))

Taken together, disjunctive statements of the form (A∨B) trigger the inference that ((¬(K(A))) ∧ (¬(K(¬(K(A)))))), or in words, that the speaker has no belief about whether A is the case or not the case. That is an ignorance implicature. In the case of (2), this will give rise to the inference that the speaker does not know whether France is a monarchy or not, which is acceptable. But in the case of (3)-(5), this mechanism of enrichment gives rise to the inference that the speaker does not know whether she herself is Germanophobic or not. That implicature, plus the common world-knowledge that one’s own attitudes are transparent, correctly predicts oddity for (3)-(5) and felicity for (2).

Comparing (5) and (3) for instance, we see how the indexical nature of the presupposition, plus the impossibility of deriving ignorance implicatures when speakers talk about their own attitudes, could dismiss the objection of contrastive behaviors under disjunctive filters. In order to control for the confounding factor of ignorance implicatures, one shall therefore test presupposition filters that trigger the right inferential mechanisms even in the presence of indexicality.

### 2.2 Subjunctive conditional filters

In order to control for ignorance implicatures in comparing slurs with other presuppositions under filters, I now compare the behavior of slurs and that of presuppositions under a presupposition filter where ignorance implicatures do not interfere. Subjunctive conditional constructions that display in the antecedent the content of a presupposition triggered in the consequent seem to constitute such a case. Compare the following conditional statements:

(8) If France was a monarchy, the monarch of France would be bald.
(9) If I were Germanophobic, then my colleagues would know that I am.

---

6 Given that each of the three alternatives asymmetrically entail the utterance, they are more informative than the utterance

7 Entailed by K(A∨B) ∧ ¬(K(A)) ∧ ¬(K(B)). Intuitively, the speaker cannot believe that A is false, as given K(A∨B), she would thus believe that B is true; but by ¬(K(B)) she does not believe that B is true. The converse entails that the speaker does not believe B to be false.
If I were Germanophobic, my colleagues would be Germanophobic too.

If I were Germanophobic, then John would be a boche.

Given that i) ignorance implicatures do not interfere in such conditionals - as shown by the acceptability of (9) for instance-, ii) the presupposition that France is a monarchy is filtered out in (8) and the presupposition that the speaker is Germanophobic is filtered out in (9) and (10), iii) the racist expressive content is not filtered out in (11), it appears that the contrast between slurs like “boche” and standard presupposition triggers restores the initial filtering problem that presuppositional views of slurs faced, even with indexical and attitudinal presuppositions à la Schlenker. But on closer inspection, one notices that again, the indexical character of the expressive presupposition, on a par with a counterfactual analysis of conditionals, could well derive the intended results. Taking R to be the relevant accessibility relation, w* the actual world, and under a dynamic strict analysis for subjunctive conditionals (von Fintel 2012), there would in fact be two alternative ways of characterizing the conditional presupposition of (11):

∀w∈R(w*)([[I am Germanophobic](c)(w)=1]→(the speaker of c is Germanophobic in w))

∀w∈R(w*)([[I am Germanophobic](c)(w)=1]→(the speaker of c is Germanophobic in w*))

That is, an utterance of (11) “If I were Germanophobic, then John would be a boche” expresses the proposition that in all (epistemically) accessible worlds where the speaker is Germanophobic, John is a German in that world (12), or alternatively, John is German in the actual world (13). As the property of being German ascribed to John is expressed through presuppositional material (under the view that “boche” carries an expressive presupposition), the difference between the two truth-conditional analyses is a crucial matter: if it is w rather than w* that is relevant for the satisfaction of the consequent, then the world variable is bound by the intensional operator, and the indexical expressive presupposition will be evaluated in the worlds that are quantified over (say, in a point of evaluation w(c)), and the anti-German sentiment will be ascribed to the utterer as she would be in this hypothetical world, not to s(c), the actual speaker of the utterance. No projection of the expressive material is thus predicted here.

But if it is w* that is relevant for the evaluation of the consequent, then the indexical expressive presupposition will be evaluated in the world of the actual context, and the anti-German sentiment will be ascribed to s(c), the utterer of (11). So Schlenker’s account can in fact predict the projection of the expressive presupposition in that case. More precisely, just like the presupposition of (8) is
conditional (Schlenker 2008), the presupposition of (11) that would be predicted under Schlenker’s analysis is either of the following:

(14)  \( \forall \omega \in \mathbb{R}(\omega^*)((\square \text{I am Germanophobic})^{c}(w) = 1) \rightarrow (s(c) \text{ is Germanophobic in } \omega) \)

(15)  \( \forall \omega \in \mathbb{R}(\omega^*)((\square \text{I am Germanophobic})^{c}(w) = 1) \rightarrow (s(c) \text{ is Germanophobic in } \omega^*) \)

In other words, an utterance of “If I were Germanophobic, then John would be a boche” is felicitous if, in all accessible worlds where the speaker is Germanophobic, the speaker is Germanophobic at that world (or alternatively, in the actual world). In the case of (14), the conditional presupposition is predicted to be trivially satisfied, and consequently, an utterance of (11) is wrongly predicted to be felicitous. But if the option to have \( \omega^* \) featuring in the computation is left open as it is the case in (15), then the conditional presupposition is not at all trivial, and imposes its non-trivial constraints on the utterance context itself: (11) is predicted to be presuppositional (in the sense that it forces hearers to accommodate the proposition that the actual speaker is prejudiced against Germans) in most contexts.

To put it in different terms, the above argument rests on the facts that, because subjunctive conditionals are intensional operators, and because Schlenker’s expressive presuppositions are indexical, the contrast between (11) and (8)-(9)-(10) is useless. If the presupposition triggered by “boche” is “indexical” in the sense that the French second person pronoun “tu” is, then it is indexical in a double way: \( i \) it is about the speaker, and \( ii \) it also imposes a condition on the world of the utterance’s context (not on the point (world) of evaluation).

Because Schlenker’s expressive presupposition is indexical, the consequent of (11) could very well presuppose that \( s(c) \) is Germanophobic in the utterance world \( \omega^* \). The world-variable is not necessarily bound by the intensional operator in the case of expressive presuppositions, and the contrast between (11) and (8) could be explained away by recognizing that the presupposition of expressives like “boche” is indexical also in the sense that their content is always to be evaluated relative to the actual world. So Schlenker’s theory does make the correct predictions here provided \( i+j \) a dynamic strict analysis for subjunctive

---

\(^8\) Roughly, the presupposition of a subjunctive conditional statement is that, in all accessible worlds, if the antecedent is true in that world, then the presupposition of the consequent is satisfied in that world (which I write, for the consequent \( q \) of an utterance in \( c \), \( \text{presup}^q(c)(w) = 1 \)). For conditional statements where the descriptive content of the antecedent precisely is the presuppositional content of the consequent, we obtain the presupposition that, in all accessible worlds, if France is a monarchy in these worlds, then France is a Monarchy in these worlds; hence the presupposition being satisfied trivially. In the case of an indexical presupposition, as one is now considering, things might be different as the satisfaction of the presupposition of the consequent might well be indexed on the actual world \( \omega^* \) rather than on the point (world) of evaluation \( w \).
conditionals and ii) a doubly indexical character of the attitudinal presupposition.

Taking stock, one cannot find a contrastive behavior under disjunctive filters because of an interfering pragmatic phenomenon of ignorance implicature, nor under subjunctive conditional filters because of the (potentially doubly) indexical character of Schlenker’s expressive presuppositions. One shall therefore test extensional contexts - that is contexts in which the world of evaluation is not affected by the conditional and is thus the same as the world of the utterance’s context - in which ignorance implicatures are controlled for. I consider below two such cases. I first go back to disjunctive filters (as they only involve extensional operators) in an imaginary and somewhat artificial context blocking ignorance implicatures. Second, I consider the case of conjunctions under negation.

2.3 Disjunctive filters without ignorance implicatures

Let us go back to disjunctive filters (as it only involves extensional operators), but this time with a context in which Grice’s maxim of quantity is suspended, resulting in the cancellation of ignorance implicatures. Imagine a world where Caucasians tend to be oppressed, marginalized, disenfranchised and so on. Imagine a game show in that world, where a participant is supposed to guess the identity of her interlocutor, hidden behind a curtain. The hidden interlocutor can give hints, like “I am the son of a baker”, or “Either I am a journalist, or I am the son of a baker”, in order to help the candidate eliminating hypotheses and eventually narrowing down her identity. In that sort of a context, “either, or” constructions do not trigger ignorance implicatures, as participants are purposefully less informative than they otherwise could be (see Fox 2014 for an earlier discussion of games in which the Maxim of Quantity is deactivated). Indeed, when the hidden interlocutor utters, “Either I am a journalist, or I am the son of a baker”, one does not infer that she does not know whether she is a journalist or not, one understands she is giving hints to the candidate rather than expressing her beliefs in maximizing informativity. As the maxim of quantity is suspended, ignorance implicatures will not be derived. Having set up this specific context will now allow us to test the projection behavior of slurs in the right kind of environment.

Context: Mary is a candidate in the game show and must guess the identity of someone hidden behind the curtain. At that stage in the game, she is hesitating between three individuals (who happen to have a daughter): Bob, an anti-Caucasian journalist whose daughter knows he is anti-Caucasian and who is anti-Caucasian too; John, an anti-Caucasian baker whose daughter does not know he is anti-Caucasian and who is not anti-Caucasian herself; and Alfred, who is notoriously not anti-Caucasian, and also has a daughter. The hidden interlocutor says: “I will give you a hint, but I shall not be too informative”:

(16) Either I don’t hate Caucasians, or my daughter knows I hate Caucasians.
(17) Either I don’t hate Caucasians, or my daughter hates Caucasians too.

Context: Similar game context, but this time with Mary hesitating between five individuals (who all have a daughter): three non-racist and two anti-Caucasian. Among the non-racists, one has a daughter who married a Caucasian, one has a daughter who did not marry a Caucasian, and we do not know about the daughter of the third. The daughter of the first racist married a Caucasian, unlike that of the other.

(18) Either I don’t hate Caucasians, or I hate Caucasians and my daughter married a Caucasian.

In (16)-(17)-(18), the racist presupposition is triggered not by a slur but by the factive “know”, the anaphoric “too”, and mere at issue content, respectively. Plus, the negation of the racist content being investigated features in the first conjuncts. Interestingly, they do not appear to convey any expressive or racist content. The three sentences seem to have been uttered as a mean to rule out some candidates (John, the anti-Caucasian baker whose daughter does not know he is, in (16)-(17) for instance), and thus express something like the disjunction “I am Alfred or I am bob”. Importantly, Mary has no grounds to draw an inference that the speaker is anti-Caucasian: she still has two options to go with, Bob and Alfred, and has no grounds to disentangle between the two. But things are different if the racist expressive content is triggered by a slur instead. Consider (19) as uttered in the same context as that of (17)-(18):

(19) !Either I don’t hate Caucasians, or my daughter married a honky.

In (19), where the anti-Caucasian content of the consequent is conveyed through the use of a slur, it seems that one can legitimately infer that the speaker is prejudiced against Caucasians. Contrary to (16)-(17)-(18), Mary does have evidence after (19) that it is Bob, the anti-Caucasian journalist, talking behind the curtain, rather than Alfred the non-racist (or than John of course). But presuppositional theories of slurs do not predict any difference between (16)-(17)-(18) on the one hand and (19) on the other hand. The presupposition of “honky” in (19) should in fact be filtered for the same reason than the “expressive” presuppositions of “know” and “too” in (16)-(17)-(18) are filtered. The fact that expressivity projects in (19) exhibits a wrong prediction of Schlenker’s and other presuppositional reductions of expressivity (in the case of slurs at least).

Now surely, given that the racist expressive content in (19) projects, the speaker must be ascribed a certain degree of hatred towards Caucasians. It is unclear then what kind of an epistemic state she could be in to make such a disjunctive statement, conveying at one and the same time a negative attitudinal state and uncertainty about her being in that state. What exactly could she have intended to communicate then? But presuppositional accounts cannot sensibly rely on such pragmatic oddity because such an oddity arrives only after the tested expressive content projects, and the mere fact that expressivity projects here is sufficient to falsify presuppositional theories. In fact, the fact that (19)
feels pragmatically odd is predicted only if the derogatory content projects, and is thus not predicted by the presuppositional approach.

In (19), it looks like the speaker intended to make a neutral disjunctive statement in order to eliminate hypotheses and ended up accidentally slipping a slur, thence revealing her true attitudes. But the fact that slipping a slur in that environment does in fact reveal her attitudes is evidence that the expressive content is not plugged where it was expected to be, for presuppositional accounts at least.

2.4 Negated conjunctions

I now turn to the second example of a non-intensional context displaying a contrast between the projective profile of slurs and that of presuppositions, one that does not involve imaginary scenarii, nor conflicting ignorance inferences.

There is a specific sort of negated conjunctions working as a presuppositional filter: the negation of a conjunction displaying in the descriptive part of the first conjunct, the presuppositional import of the second conjunct. First, note that such conjunctions are not presuppositional (Karttunen 1973):

(20) France is a monarchy and the monarch of France is bald.

(20) does not presuppose that France is a monarchy. Indeed, although the consequent alone does carry the presupposition that France is a monarchy, adding precisely that content as the first conjunct has the effect of restricting the context set of the evaluation of the consequent precisely to those worlds in which the presupposition of the consequent is already satisfied. No further restriction is needed (in a dynamic framework, see again Schlenker 2008), that is, (20) is not presuppositional. Now, of course, an utterance like (20) will still convey the false information that France is a monarchy, but that is only because of its descriptive material; it is said that France is a monarchy. One can therefore safely construct its negated alternative:

(21) It’s completely false that France is a monarchy and that the monarch of France is bald.\(^9\)

And as expected, the result is neither presuppositional nor does it convey the false information that France is a monarchy. Now consider the same constructions involving Germanophobic content:

(22) I am Germanophobic, and my colleagues do not know it.

In (22), the context set for the evaluation of the second conjunct is restricted to precisely those worlds which satisfy the presupposition, so that (22) is not...

---

\(^9\) I use the unnatural “it’s completely false that” form for negation in order to avoid potential complications that could arise because of the ubiquitous metalinguistic readings of standard negation, (provided that phrase actually blocks metalinguistic construals).
predicted to be presuppositional. An utterance like (22) will still be very offensive in virtually any context, but again that is only because of its descriptive material: it is said that the speaker despises German people. And since (22) expresses anti-German sentiment because of its descriptive material, one can safely construct its negated alternative and expect the result to be neutral:

\[
(23) \quad \text{It's completely false that I am Germanophobic and that my colleagues do not know it.}
\]

And indeed, (23) does not seem to commit the speaker to anti-German sentiments. So far so good. I note here that for puzzling reasons that ought to be clarified independently, (23) and other such constructions appear not to trigger ignorance implicatures, although they are equivalent to some disjunctive filters such as our earlier (2)-(5)\(^{10}\). (23) is indeed formally equivalent to the odd (3), but where (3) resulted in oddity because of the intervention of ignorance inferences, (23) does not. This difference will be helpful.

Consider the following utterances, constructed by equivalence on the model of (2)-(5):

\[
\begin{align*}
(24) & \quad \text{It's completely false that France is a monarchy and that the monarch is hairy.} \\
(25) & \quad \text{It's completely false that I am Germanophobic and that my colleagues (do not) know it.} \\
(26) & \quad \text{It's completely false that I am Germanophobic and that my colleagues are Germanophobic too.} \\
(27) & \quad \text{It's completely false that I am Germanophobic and that John is (not) a boche.}
\end{align*}
\]

Be it in the current context or in an arbitrary one, a speaker uttering (27) would still be seen as displaying her negative attitudes towards German people. That is inexplicable for a presuppositional view of slurs, as shown by the felicity and neutrality of (25)-(26). For the reasons evoked just above, (27) is not presuppositional. Furthermore, one cannot draw any sort of racist expressive content from its descriptive layer. Presuppositional views of slurs à la Schlenker thus have nowhere to locate the source of that anti-German content leaking out of (27). Presuppositional accounts of slurs make the prediction that (27) is neutral; we observe it isn’t. I conclude, from the case study of slurs, that expressive content is not presuppositional.

If slurs do not convey their racist expressive content via a presuppositional mechanism, then the way is open to Pott’s use-conditional analysis of expressivity (Potts 2007), or maybe to more strictly pragmatic accounts (Nunberg 2016). The above data could also constitute evidence that the class of projective content is broader, and hence non-reducible, to that of presuppositions. That could

\(^{10}\) By De Morgan and classical rules for double negation, \((\neg A \lor \neg B)\) is equivalent to \(\neg (A \land \neg B)\).

195
contribute to falsify different attempts at reducing projective content in general to the narrower class of presuppositional content (see Roberts et al. 2010 or Tonhauser et al. 2013 for discussion).

3 References

Roberts et al. 2010 or Tonhauser et al. 2013 for discussion.

3 References


• Kaplan, D. (2001), The Meaning or Ouch andOops (Explorations in the theory of Meaning as Use), Draft #3, ns., UCLA.
• Magri, B. (2009), A theory of individual-level predicates based on blind mandatory scalar implicatures, Natural Language Semantics 17:245297.
• McCready, E. (2010), Varieties of conventional implicature, Semantics and Pragmatics, 3(8), 1-57.
• Strawson, P. F. (1964), Identifying reference and truth values, Theoria 30:96118.
• Spector, B. (2003), Scalar implicatures: Exhaustivity and Gricean reasoning, In B. ten Cate (Ed.), Proceedings of the Eighth ESSLLI Student Session.
Exploring Compositionality of Estonian Particle Verbs

Eleri Aedmaa
Institute of Estonian and General Linguistics, University of Tartu
eleri.aedmaa@ut.ee

Abstract. Traditionally Estonian particle verbs are divided into two classes subject to the formation of their meaning – compositional and non-compositional. First, we compare several lexical association measures for the automatic detection of the degree of compositionality of particle verbs. Second, we test the power of distributional semantics to rank Estonian particle verbs based on their degree of compositionality. The results are evaluated against the test data set – the first human-rated compositional ranking for Estonian particle verbs.

1 Introduction

Multi-word expression (MWE in what follows) is recurrent combination of two or more words (not necessarily contiguous) that form (semi-)fixed lexical or syntactic unit [35]. An Estonian particle verb (PV in what follows) is a MWE that consists of a verb and a verbal particle, e.g. in (1) the verbal component of PV maha vötma 'to take down' is võeti 'was taken' and the particle is maha 'down'. Similarly to other MWEs, PVs may be more or less opaque in compliance with the individual meanings of their components. According to [31], repeated later in [12,13], the Estonian PV can be classified as compositional or idiomatic depending on how its meaning is composed from the meanings of its components. The components of the compositional PV take their literal meaning, e.g. in (1) the meaning of the PV maha vötma 'to take down' is a composition of the meanings of its components maha 'down' and vötma 'to take'. The meaning of non-compositional PV is idiosyncratic and can not be inferred from the literal meanings of the verb and the particle, e.g. (2) the meaning of the PV maha võeti is 'to lose weight'. To the best of our knowledge, the view that MWEs form a continuum between entirely compositional and entirely non-compositional expressions [25], has not been broadly adopted by Estonian theoretical linguists.

(1) Lava võeti kohe pärast kontsert-i maha.
    stage take-PST,PASS right after concert-PART.SG down
    ‘The stage was taken down right after the concert.’

(2) Ta on kümme kilo maha võt-nud.
    he be.3SG,PRS ten kilogram down take-PST,PTCP
    ‘He has lost 10 kilograms.’
Estonian has a relatively free word order and the syntax allows discontinuous realization with a large distance between the components of the PV, as in the examples (1–2). In addition, the order of a particle and a verb may vary, e.g. in (1) the particle maha follows the verb võtma, in (2) the particle maha lies between auxiliary verb olema ‘to be’ and the verb võtma.

Series of studies have been conducted in order to extract and identify MWEs in various languages [10, 20, 27], including Estonian where the main focus has been on multi-word verbs [17, 18, 36]. For example, Aedmaa [1] applied various statistical association measures (AMs in what follows), i.e. mathematical models that interpret word co-occurrence frequencies in a given text span and compute an association score for any pair of words to indicate the degree of association between the two words, to automatically detect PVs from corpora. Less work has been done regarding predicting the compositionality of PVs. Most studies have been carried out on English [3, 5, 25] and German [7, 21]. One of the first studies was by Lin [23] who used similarity measure to identify non-compositional phrases. Its most influential contribution for studies to come was the inference that non-compositional expressions have different distributional features than compositional ones.

Distributional semantics has been applied to detect the compositionality of MWEs, e.g. [19, 33, 34], but also to rank them according to their multiwordness, e.g. [6, 30]. The main approach to these tasks is to use measures (e.g. cosine similarity or Euclidean distance) that express the similarity between the vector of a MWE and the vectors of the constituent words of the MWE. Created models are usually evaluated against gold standards based on human judgement. For example, for German there are three compositionality ratings available [6]. We introduce the first compositionality ranking for Estonian MWEs.

This is an opening study to (automatically) analyse an exhaustive list of Estonian PVs based on the formation of their meaning. In Section 2, we introduce annotated corpus and compositionality rating for PVs. Section 3 describes experiments and results. We test if AMs can identify compositional and/or non-compositional PVs and then we run experiments to place PVs on a continuum according to their degree of compositionality applying first AMs, and then distributional semantics. Results of the experiments are discussed in Section 4. We conclude and outline future directions in Section 5.

2 Dataset and Compositinality Rating

In [1] the applied AMs automatically extracted 1676 PVs from morphologically-analysed and disambiguated newspaper texts (170 million tokens) of Estonian Reference Corpus. For the human-rated ranking we took a list of these PVs and removed infrequent ones (i.e. occurred less than 9 times in the corpus) from this set, and randomly sampled 193 PVs from different frequency classes (we sorted PVs according to their frequency and added every 9th PV to the dataset). We ensured that 20 most frequent PVs are also included in the dataset, so we collected

\(^1\) http://www.cl.ut.ee/korpused/segakorpus/
rates for 211 PVs. We asked 110 judges to evaluate to what extent the meaning of a PV agrees with the meanings of its components. We collected compositionality ratings via Qualtrics survey link, only Estonian native speakers were asked to participate. Each annotator evaluated 21 PVs on a scale from 1 (non-compositional) to 5 (compositional), with the 6th option I don’t know included. We decided to use a scale with an odd number of categories because Estonian PVs are often ambiguous between compositional and non-compositional readings, and we assumed that in this case humans choose not to specify if the PV is compositional or non-compositional. Each PV was evaluated at least 10 times.

The 6th option I don’t know was asked to use only when the evaluation is very difficult or even impossible. For 40 PVs, the annotation I don’t know was assigned once, for 11 PVs twice, and for 3 PVs thrice. These PVs are not included to the final dataset. The reasons why it was impossible for judge(s) to rate the compositionality of some PVs should be studied more thoroughly, but it could be due to the annotator’s subjective opinion or because of the ambiguity of one or both components of the PV. In fact, most of the PVs that got at least one I don’t know rate have polysemous components. For example, the verbal component ajama of the PV maha ajama ‘knock off’, ‘shave off’, ‘push off’ has 16 meanings presented in the Estonian Explanatory Dictionary. Particle maha ‘down’ indicates prototypically direction with the meaning ‘down’, but it also can express perfectivity when its meaning is ‘off’ or ‘out’.

To evaluate the quality of a dataset, measures that indicate agreement among different annotators should be employed. The main measure for calculating agreement is the (Cohen’s or Fleiss’) kappa coefficient. Due to lack of human resources different annotators rated different PVs and it is not possible to calculate agreement between annotators for our compositionality dataset. However, we follow [9] and calculate the standard deviation (σ) for each PV (except 54 PVs that were at least annotated with I don’t know). High σ value refers to low agreement among annotators, while low σ value shows that annotators gave similar rates.

15 PVs have a standard deviation higher than 1.50, following [29], we used it as a threshold to note that agreement among annotators is low enough to treat these PVs as outliers. The high disagreement seems to be caused by the polysemous particles, because most of the verbal components are monosemous, e.g. laadima ‘to load’, reastuma ‘to align’, pakkima ‘to pack’, üürima ‘to rent, loosima ‘to draw lots’. Low agreement among judges might be also caused by the different background of annotators – regrettably we did not ask judges to specify whether they have studied linguistics or not. The final dataset consists of 142 PVs.

---

2 www.qualtrics.com
3 See Section 1 for details.
4 http://www.eki.ee/dict/ekss/
5 The dataset is available at https://github.com/eleriaedmaa/compositionality.
3 Experiments and Results

3.1 Dividing PVs into Fixed Classes

Findings of Aedmaa [1] hint that AMs can be used for the identification of different types of PVs. In particular they showed that several AMs – t-score, mutual information (MI) [8], $X^2$ [24], log-likelihood [11], minimum sensitivity (MS) [28], and co-occurrence frequency – are successful for the Estonian PV extraction task in general. Hence, we hypothesize that AMs differentiate fully compositional and fully non-compositional PVs from other PVs, similarly to [14, 15], which used AMs to identify expressions that may potentially have idiomatic meanings. As the previous studies have claimed that semantic compositionality is continuous [2], we do not expect AMs to divide PVs into two classes. We test if a certain AM extracts one or both types of PVs with high recall such that that we can use this method for the identification of very compositional and non-compositional PVs. The result of this experiment might also support or refute the aforementioned theoretical approach about the classification of Estonian PVs.

We calculate the proportion of compositional and non-compositional PVs each AM extracts from the corpora. First, for the AM rankings, five different association scores for all PVs are computed. A higher association score indicates stronger association between the components of the PV. Second, to calculate the percentages of compositional or non-compositional PVs each AM detected, PVs are ranked according to the association score value and evaluated against the division of PVs based on human judgement. For the assessment we average the rates of compositionality given by judges for 157 PVs (we excluded PVs that were at least once evaluated as I don’t know) and form two groups\(^6\) – compositional (PVs with average rating larger than 3.5), and non-compositional (PVs with average rating less than 2.5). We focus on the two extreme cases – fully compositional and fully non-compositional PVs – so analyses of the other PVs are beyond the scope of the paper. As there were 92 compositional and 8 non-compositional PVs in our dataset, the formation of meaning of the Estonian PV is more likely compositional than non-compositional.

\(^6\) The division is rough and we use it only for the evaluation.

Table 1. Recall values of 50 highest-ranked PVs according to their type and correlation between rankings of AM and human compositionality ratings.

<table>
<thead>
<tr>
<th>AM</th>
<th>compositional</th>
<th>non-compositional</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-score</td>
<td>32.6%</td>
<td>37.5%</td>
<td>0.02</td>
</tr>
<tr>
<td>MI</td>
<td>29.3%</td>
<td>37.5%</td>
<td>-0.07</td>
</tr>
<tr>
<td>$X^2$</td>
<td>30.4%</td>
<td>37.5%</td>
<td>-0.03</td>
</tr>
<tr>
<td>log-likelihood</td>
<td>32.6%</td>
<td>37.5%</td>
<td>-0.01</td>
</tr>
<tr>
<td>MS</td>
<td>32.6%</td>
<td>37.5%</td>
<td>-0.03</td>
</tr>
<tr>
<td>frequency</td>
<td>37.0%</td>
<td>37.5%</td>
<td>-0.00</td>
</tr>
</tbody>
</table>
The percentages given in Table 1 express how many compositional and non-compositional PVs applied AMs extracted among 50 PVs that got the highest association scores of the corresponding AM. For example, there were 32.6% of all compositional PVs and 37.5% of all non-compositional PVs among top 50 PVs that get highest MS score. All AMs were able to extract only 3 non-compositional PVs. Frequency detected slightly more compositional PVs than other AMs. There were fully compositional and fully non-compositional PVs among 50 highest ranked PVs for all AMs, so none of the AMs extract well only compositional or non-compositional PVs. Moreover, recall values are relatively low and no AM performs significantly well. Hence, compositional or non-compositional Estonian PVs are not automatically identifiable using statistical AMs.

### 3.2 Detecting Degree of Compositionality of PVs

Compositionality of MWE is an consequence relationship between the whole and its various parts [4]. Despite the fact that AMs are not traditionally used for the compositionality tasks, we still calculated the Spearman’s correlation coefficient (\( \rho \)) for each AM in order to test if AMs are suitable for detecting the degree of compositionality of PVs. For this purpose the ranking based on the average values of human judgement was created and compared to the rankings of AMs. The \( \rho \) values in the Table 1 indicate that there is no linear relationship between the rankings. Hence, it is not possible to automatically predict the degree of compositionality of PVs using statistical AMs.

In order to use a distributional semantics model for the detection of compositionality degree of PVs, we adopt the idea from Bott and Schulte im Walde [6] who hypothesized that the more similar the contexts of a PV and a corresponding (base) verb are more similar their meanings tend to be. In addition, according to [12] the verbal component of the PV is consider to be substantial meaningful component of the PV, while the particle expresses direction, perfectivity, state or modality. A context of a compositional PV is less influenced by the particle than a context of a non-compositional PV. For example, in (3) and (4) the contexts of the compositional PV eemale tõukama ‘to push away’ and the verb tõukama ‘to push’ are very similar, because both sentences are describing the act of pushing, though the object case differs.

(3) Naine tõuka-s mehe eemale ja ta istu-s maha. woman push-3SG.PST man.SG.GEN away and he sit-3SG.PST down
   ‘Woman pushed the man away and he sat down.’

(4) Naine tõuka-s mees-t ja ta kukku-s maha. woman push-3SG,PST man-SG.PART and he fall-3SG.PST down.
   ‘Woman pushed the man and he fell down.’

At the same time the non-compositional PV vahele kukkuma ‘to get caught’ is used to describe some kind of cheating (using doping, stealing, criminal activity,
etc), but *kukkuma* 'to fall' usually occurs with words expressing the matters (subject, place, time, reason, consequence) of falling down.

In this experiment the similarity between the PV and the verb is expressed by their cosine similarity (CS). Higher value of CS indicates higher degree of compositionality of PV. For the embeddings, we use word2vec’s Continuous Bag-of-Words model (CBOW) [26] with 200 dimensions and window size 10, and Gensim toolkit for Python. The model was trained on the same but lemmatised newspaper texts as introduced in the Section 2. The employed model is able to find the CS value only for continuous PVs, i.e. PVs with adjacent components. We used non-compositional embeddings – as we concatenated components of PVs prior to training word embeddings, PVs are treated as a single unit. Thus, when we computed embeddings, the vector of a base verb does not include the context of the PV and represents the meaning it possesses by itself. In the future, other (and more sophisticated) models, such as suggested by Hashimoto and Tsuruoka [16], should be tested.

We rank the PVs according to their CS values. This continuum from compositional to non-compositional PVs is compared to the ranking based on human judgement. Because of the reasons mentioned above we investigate only the similarity between a PV and its base verb when calculating the degree of compositionality of the PV. Humans at the same time also considered the contribution of the meaning of the particle. This difference might have an impact on the results, and it should be investigated in the future. Also, it is worthy to note here that when evaluating the similarity between a PV and its components, human judgement is based on the prototypical meaning of the base verb which might not be the most frequent in the corpus. So it might be the case that the model and a human evaluate different meanings of the same verb.

Table 2. Correlation between two rankings for different frequency sets of PVs.

<table>
<thead>
<tr>
<th>set of PVs</th>
<th>frequency</th>
<th>$\rho$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>9–28,435</td>
<td>0.27</td>
<td>0.01</td>
</tr>
<tr>
<td>without frequent</td>
<td>$&lt;10,000$</td>
<td>0.23</td>
<td>0.01</td>
</tr>
<tr>
<td>without infrequent</td>
<td>$&gt;23$</td>
<td>0.18</td>
<td>0.05</td>
</tr>
<tr>
<td>without frequent and infrequent</td>
<td>23–10,000</td>
<td>0.16</td>
<td>0.05</td>
</tr>
<tr>
<td>frequent</td>
<td>$&gt;10,000$</td>
<td>0.09</td>
<td>0.69</td>
</tr>
<tr>
<td>infrequent</td>
<td>$&lt;23$</td>
<td>0.29</td>
<td>0.21</td>
</tr>
</tbody>
</table>

From Table 2 it can be seen that among all 142 PVs the correlation coefficient is 0.27 and it is statistically significant at the level of 0.05. The correlation value decreases when 20 most infrequent (i.e. occurring less than 23 times) and/or 20 most frequent (i.e. occurring more than 10,000 times) PVs are removed from the data. Among frequent PVs the correlation is lower than among infrequent ones, but the correlation for frequent and infrequent PVs is not statistically significant.

https://radimrehurek.com/gensim/
(this fact might be caused by the small size of the sample). Although the further reasons of the poor correlation between the rankings are discussed in the next section, we can say that it is difficult to predict the degree of compositionality for frequent PVs. Bott and Schulte im Walde [7] concluded the same about German PVs.

4 Discussion

In the following, we consider the matters that might have influenced the results. Table 3 introduces most compositional and non-compositional PVs according to the CS value and gives their frequency and their average rate of compositionality according to human judges (HJ).

<table>
<thead>
<tr>
<th>CS</th>
<th>particle verb</th>
<th>in English</th>
<th>frequency</th>
<th>HJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.44</td>
<td>maha 'out' műüma 'to sell'</td>
<td>to sell</td>
<td>8435</td>
<td>2.5</td>
</tr>
<tr>
<td>0.43</td>
<td>tagasi 'back' minema 'to go'</td>
<td>to go back</td>
<td>4522</td>
<td>4.6</td>
</tr>
<tr>
<td>0.38</td>
<td>üle 'ever' küsíma 'to ask'</td>
<td>to ask again</td>
<td>481</td>
<td>2.8</td>
</tr>
<tr>
<td>0.37</td>
<td>vikja 'out' kuulutama 'to announce'</td>
<td>to announce</td>
<td>13196</td>
<td>3.2</td>
</tr>
<tr>
<td>0.37</td>
<td>vastu 'back' küsíma 'to ask'</td>
<td>to ask as a reply</td>
<td>517</td>
<td>3.9</td>
</tr>
<tr>
<td>-0.07</td>
<td>vahele 'between' kukkuma 'to fall'</td>
<td>to get caught</td>
<td>15</td>
<td>2.2</td>
</tr>
<tr>
<td>-0.09</td>
<td>vülja 'out' saagíma 'to saw'</td>
<td>to cut out with saw</td>
<td>59</td>
<td>4.4</td>
</tr>
<tr>
<td>-0.10</td>
<td>vülja 'out' pilduma 'to throw'</td>
<td>to throw out</td>
<td>19</td>
<td>3.6</td>
</tr>
<tr>
<td>-0.10</td>
<td>pael 'on' tungíma 'to force'</td>
<td>to attack</td>
<td>175</td>
<td>2.8</td>
</tr>
<tr>
<td>-0.27</td>
<td>kokku 'together' kiskuma 'to maul'</td>
<td>to wrench</td>
<td>17</td>
<td>4.3</td>
</tr>
</tbody>
</table>

CS values indicate that the most compositional PV is maha műüma 'to sell' and the most non-compositional is kokku kiskuma 'to wrench'. The particle maha in the PV maha műüma expresses perfectivity and the PV has a very similar meaning to its verbal component műüma 'to sell'. Hence, we can consider this PV highly compositional. The rating of human judgement is low because particle maha is prototypically used in directional meaning 'down', which probably makes it difficult to rate the compositionality of this PV.

Frequency has clear influence on results – it is visible that PVs with high value of CS are more frequent than the PVs with low CS value. This observation gives evidence that the frequency of the PV may influence the calculation of CS. The correlation between the frequency of PV (and its components) and the CS value needs to be studied in the future. Tagasi minema 'to go back' is the only PV which appears among 10 most compositional PVs in both rankings. Although this PV has multiple meanings, all of them are compositional.

Since we are able to find the CS value only for PVs which components occur next to each other in the data, the frequency of some PVs becomes very low. For example, components of 19 of the 20 most infrequent PVs occur less than 5
times adjacent to each other in the whole corpus, but the components of frequent PVs occur side-by-side frequently enough.

In addition, Figure 1 that illustrates the correlation between the values of the CS and the average of human judgements, shows clearly that the 20 most infrequent PVs (denoted as squares) have lower CS values than frequent PVs (denoted as triangles). We can hypothesize that the used model tends to calculate higher CS values for frequent PVs. If it is the case, then the poor correlation among frequent PVs can be partly explained by the fact that the frequent PVs are not as compositional as the value of CS indicates. As there are highly compositional and highly non-compositional PVs among frequent PVs, the CS values do not sufficiently indicate the degree of compositionality of frequent PVs. Compared to infrequent PVs, the verbal component of frequent PVs are very polysemous, i.e. minema 'to go', tulema 'to come', pidama 'must', 'have', 'to keep'. This complicates the evaluation process for humans and the model does not consider several different meanings of a verb. So, as both rankings are influenced by the frequency, the high correlation of rankings among frequent PVs would be random. As the verbal components of the infrequent PVs are less ambiguous, e.g. voltima 'to fold', maksma 'to pay', kukkuma 'to fall', then the rankings are more solid and the correlation is higher.

![Fig. 1](image.png)

Fig. 1. The correlation between values of CS, and the averaged ratings of human judgement among frequent, infrequent and all PVs.

Table 4 introduces most compositional and non-compositional PVs according to their average rate of compositionality based on human evaluations (HJ),
their frequency and standard deviation which reflects the agreement among annotators. According to human judgement the most compositional Estonian PV is *eemale tõukama* ’to push away’ and the most non-compositional are *välja nägema* ’to appear’, ’to see outside’ and *vastu põrutama* ’to talk back’.

Table 4. The most compositional and non-compositional PVs according to human judgement.

<table>
<thead>
<tr>
<th>HJ</th>
<th>particle verb</th>
<th>in English</th>
<th>frequency</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.7</td>
<td>eemale ’away’ tõukama ’to push’</td>
<td>to push away</td>
<td>179</td>
<td>0.78</td>
</tr>
<tr>
<td>4.6</td>
<td>ette ’ahead’ joudma ’to reach’</td>
<td>to get ahead</td>
<td>1542</td>
<td>0.50</td>
</tr>
<tr>
<td>4.6</td>
<td>tagasi ’back’ minema ’to go’</td>
<td>to go back</td>
<td>4522</td>
<td>0.79</td>
</tr>
<tr>
<td>4.5</td>
<td>jaelle ’after’ kihutama ’to race’</td>
<td>to give chase</td>
<td>52</td>
<td>0.52</td>
</tr>
<tr>
<td>4.5</td>
<td>sisse ’in’ kutsuma ’to invite’</td>
<td>to call in</td>
<td>171</td>
<td>1.13</td>
</tr>
<tr>
<td>2.2</td>
<td>vahele ’between’ kukkuma ’to get caught’</td>
<td>to get caught</td>
<td>15</td>
<td>1.17</td>
</tr>
<tr>
<td>2.2</td>
<td>taga ’back’ kihutama ’to race’</td>
<td>to encourage</td>
<td>22</td>
<td>1.08</td>
</tr>
<tr>
<td>2.0</td>
<td>üles ’up’ kloppima ’to fluff’</td>
<td>to fluff up</td>
<td>31</td>
<td>0.74</td>
</tr>
<tr>
<td>1.9</td>
<td>välja ’out’ nägema ’to see’</td>
<td>to appear, to see outside</td>
<td>11105</td>
<td>0.74</td>
</tr>
<tr>
<td>1.9</td>
<td>vastu ’back’ põrutama ’to knock’</td>
<td>to talk back</td>
<td>112</td>
<td>1.00</td>
</tr>
</tbody>
</table>

It is important to bear in mind that frequency does not influence only the automatic prediction of the degree of compositionality. It is a well-known fact that most frequent words are often polysemous, which makes the degree of compositionality difficult to evaluate. For example, *välja nägema* ’to appear’, ’to see outside’ is a frequent PV in the corpus and it has literal and non-literal meanings. Additionally, both components are frequent and have multiple meanings. All these circumstances together make difficult to evaluate its compositionality. The same conclusion is valid for infrequent *vahele kukkuma* ’to get caught’, because its frequent components have several meanings.

Although it is more difficult to predict the compositionality degree for frequent PVs, the frequency essentially influences our results. Thus, it is necessary to employ models that are more successfully used to tackle the problem of data sparsity, e.g. compositional embeddings. We also plan to carry out comparison with the similar phenomenon in other languages and ascertain linguistic features of Estonian PVs that must be taken into consideration when detecting their compositionality.

5 Conclusions

We did not succeed to automatically predict the degree of compositionality of PVs using statistical AMs, but we were able to automatically rank Estonian PVs based on their degree of compositionality. However these rankings correlate poorly with human-rated compositional ranking of PVs and the applied model
needs to be developed in order to improve the quality of detecting the compositionality of Estonian MWEs in general. In addition to aforementioned plans two goals need to be addressed in the future: 1) to take into account PVs the components of which are not adjacent to each other in the sentences, and 2) how to detect different meanings of polysemous components of PVs.

6 Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This research was supported by the European Union through European Regional Development Fund and Center of Excellence in Estonian Studies under the Estonian Ministry of Education and Research project Computational models for Estonian (IUT20-56).

References

10. Daille, B.: Study and implementation of combined techniques for automatic extraction of terminology. The balancing act: Combining symbolic and statistical approaches to language, 1:49–66 (1996)


The Effect of Poor Coreference Resolution on Document Understanding

Maximilian Droog-Hayes

School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, United Kingdom
m.droog-hayes@qmul.ac.uk

Abstract. For the majority of documents, human readers can understand which entities are being referred to with little difficulty, which is an important step in the interpretation of a text. Coreference resolution is a commonly used preprocessing step in tasks such as automatic summarization to identify references to the same entity. The metrics used to evaluate coreference resolution systems indicate they are at an acceptable level of performance, however previous work has suggested that they do not perform well enough for summarization. This paper describes a preliminary experiment designed to investigate the degree to which such tools are useful for automatic summarization. The results obtained suggest that coreference systems somewhat contribute useful information, but that intrinsic evaluation metrics give misleading results: further extrinsic evaluation of coreference resolution systems is therefore needed in order to fully assess their usefulness for practical tasks in computational linguistics.

Keywords: Coreference Resolution · Evaluation · Summarization · Document Understanding

1 Introduction

Many tasks in natural language processing such as summarization and question answering require a level of understanding of the text being considered. One issue encountered in such tasks is that entities in a text often have co-referring mentions – they are referred to by different noun phrases. Coreference resolution is the task of grouping together the expressions that refer to the same entities. This is far easier for human readers than for automatic systems, as humans can often determine the coreferentiality of two mentions from just the mentions and some additional context, when it is required [15]. Consider:

Jon went to the shop. He bought a drink.

In the above, He unambiguously refers to Jon but the situation becomes more complex in longer texts as more characters are introduced and mentions may occur further away from their referent.
Our overall research programme is one of abstractive summarization, a task which requires a greater level of document understanding than extractive summarization, which depends on shallow statistical techniques [7]. End-to-end summarization contains multiple processing steps, each potentially introducing errors; and for abstractive summarization, high quality coreference information may be crucial. However, its exact contribution is unclear, as is the level of accuracy required. In addition, the evaluation of summaries can be expensive and highly subjective; to assess the contribution of coreference information we need a more objective measure, but while intrinsic coreference metrics are available, it is unclear how well they relate to overall task performance.

In this study, then, we examine the contribution of coreference resolution to semantic understanding by testing the extent to which it aids people in a closely related task: answering questions about short stories. These questions were designed to cover the critical aspects of the stories: those that must be understood in order to create a summary. The aim of this work is not to find the best coreference resolution system, but to see whether systems can be used in tasks which require document understanding without contributing significant inaccuracy; and to examine the utility of intrinsic coreference evaluation metrics.

In section 2 we discuss previous literature which describes the importance of coreference resolution as a preprocessing step in tasks such as summarization, as well as discussing the intrinsic evaluation metrics used in this work. Section 3 explains our experimental set-up. Sections 4 and 5 discuss the intrinsic and extrinsic evaluation results on our dataset. We use the results of our experiment to suggest whether automatic coreference resolution can be used for tasks such as summarization, and whether the scores obtained on standard coreference resolution evaluation metrics truly indicate their extrinsic performance.

## 2 Related Work

Coreference resolution has previously been considered an important preprocessing step for higher level tasks in Natural Language Processing. Work relating to both of the systems tested here alludes to its importance. Lee et al. [8] write that coreference resolution is important for tasks like summarization, question answering, and information extraction. Bengston and Roth also mention the importance of coreference as a prerequisite to a variety of tasks such as textual entailment and information extraction [3]. Several of the skills that [19] proposes as necessary for comprehensive question answering also cover different aspects of coreference resolution and indicate its importance.

The benefit of coreference resolution to summarization has been previously identified by Steinberger et al. [16] in their Latent Semantic Analysis based approach to summarization. In their work the authors use anaphoric information for both giving additional information to the latent semantic representation of a given document which is used to produce a summary, and to check that anaphoric expressions in the extracted summary can still be interpreted, in case their antecedent has been omitted from the final output. The authors
find that adding anaphoric information leads to a performance improvement in summarization, even when this is imperfect automatically obtained data rather than ideal manual coreference resolution. While this examines the improvement gained from adding anaphoric information to an extractive summarizer against a fixed reference, here we will investigate the extent to which coreference resolution is necessary for understanding a narrative and whether automatic methods are accurate enough to be used in the creation of summaries.

Many summarization systems use coreference resolution as a preprocessing step, in its most basic form this simply involves determining the most important characters by observing the entities that have been assigned the greatest number of mentions. However Genest and Lapalme [7] have suggested that freely available coreference resolution systems are not even remotely close to the required level of performance. They claim that generating summaries containing false information is a very serious problem, but much less prevalent in extractive summaries than abstractive summaries, only occurring in extractive systems when anaphoric words are incorrectly resolved. Caselli et al. [6] found that preprocessing steps such as named entity recognition and classification contributed errors to their higher level task of constructing event timelines, and that the performance of these systems is less than is indicated by results on benchmark datasets.

The two systems are compared here using three intrinsic coreference resolution evaluation metrics: MUC [18], \( B^3 \) [1] and the entity based CEAF variant, CEAF\(_e\) [9]. Cai and Strube [4] provide a detailed description of all three metrics along with their relative strengths and weaknesses.

The MUC metric handles the scoring of coreference resolution systems by looking at the number of links that would need to be added or removed from a system response in order to replicate the set of gold standard entities and their mentions. As this metric scores systems in terms of links between mentions it should not be used on datasets with singleton entities, entities which only have a single mention in a text. This is because singleton entities do not have multiple mentions to link, so it does not credit systems for separating singleton entities from other coreference chains.

\( B^3 \) takes singletons into account, by calculating precision and recall scores for each mention and then averaging over all mentions. Unlike MUC this metric does not consider all types of errors to be equal, and some are penalized more harshly than others. \( B^3 \) does however assume that both the gold standard and the system response are comprised of matching sets of twinned elements [17], which is not the case when the system is automatically identifying mentions as well as clustering them into coreferent entities.

CEAF uses the best alignment of system to gold responses in order to calculate precision and recall in terms of either mentions or entities. The reason for this is that [9] finds the results of both MUC and \( B^3 \) to be counter-intuitive as entities are used more than once in their calculation. However CEAF weights the alignment of all entities equally, regardless of the size of the coreference chain.
3 Approach

3.1 Materials

The two automatic coreference resolution systems used in this study were Stanford’s CoreNLP [8, 10] and the Illinois coreference resolution system [12]. The CoreNLP coreference resolution system is a multi-pass sieve of deterministic coreference models, applied one by one from highest precision to lowest precision. The Illinois coreference resolution system is a pairwise classification model, deciding if mentions belong to the same equivalence class, trained on the Automatic Content Extraction (ACE) datasets.

A dataset of 10 stories was assembled from a collection of Aesop’s fables. These short self-contained stories were chosen as opposed to segmenting a longer text, as separating character mentions could unfairly penalize the coreference resolution systems. As many of Aesop’s fables contain anthropomorphic animals, half of the selected stories specifically involve only human characters. This distinction is present to show that the automatic coreference resolution systems being tested are not disadvantaged by this type of data; they are often trained on corpora from sources such as news articles, magazines and web blogs [12], which do not usually contain animals with human characteristics. The stories involving only humans contained 96 noun mentions and those involving animals contained 118 noun mentions.

A first set of participants were asked to produce questions and corresponding answers regarding the summary-worthy aspects of each story. The answers given in this step were taken as the ground truth values that the results of the next step should be compared to, where a second set of participants answered these questions with specific coreference information. This question answering task was performed on graphical representations of the text as opposed to the original plain text. By showing participants a fixed meaning representation of the text, this alleviates the possibility of participants giving a range of answers to the questions due to differing interpretations of a story’s meaning. Abstract Meaning Representation (AMR) [2] was used as the semantic representation language for this task. Tree-like visualizations of the texts’ AMR parses were the main component of the graphical representation, along with PropBank [11] edge labels and coreference information.

3.2 Method

The questions and answers given by the first set of participants were combined for use in the rest of the study according to the following rules: Questions that could not be explicitly answered from the text, such as most why questions, were removed to avoid any ambiguity as to the correct answer. Repeating occurrences of the same question were removed (different participants very often provided near identical questions), leading to a set of 42 questions. Finally, the wording of questions was modified to match the source text as closely as possible. As an example:
Text excerpt: The wolf chased the rabbit.
Respondent question: Who ran after the rabbit?

For the above, the question text would be modified to *Who chased the rabbit?* This was carried out to make the questions as unambiguous as possible, given the graphical representation presented to the second set of participants.

Manual AMR parses for each text were then created for the construction of tree-like graphical representations of each story. These differed slightly from standard AMR parses in that intra-sentential mentions of a character were not resolved to the same entity, which would act as manual coreference resolution. It was ensured that exactly the same number of character mentions occurred in the parses as in the original source text of each story, so that the outputs of automatic coreference resolution systems could be correctly annotated on the visualizations. For AMR parses of sentences containing coordinating conjunctions, the subject of the first clause was explicitly linked to the second clause. For humans reading plain text, it is self-evident that the subject is the same in both clauses of such sentences, but this may not be the case when reading AMR parses and so the link is added to ensure that related questions can be answered without difficulty. This was also performed as these conjunctions join two verb phrases with only a single subject, so there is no second noun phrase present for automatic coreference resolution systems to try and detect, let alone correctly resolve.

The graphical representation of a given story consisted of a force directed graph with each sentence’s AMR parse. AMR parses do not show the PropBank frame roles that describe the links between two AMR entities. These were parsed onto the visualizations so that hovering on a link between two nodes displayed a description of the link type. Consider the graph fragment in Fig. 1 for the text *The wolf chased the rabbit.* From this alone it is not possible to tell whether the wolf is chasing the rabbit or the rabbit is chasing the wolf. Adding on the relevant PropBank roles *follower* and *thing followed* makes this unambiguous.

![Graphical representation](image)

**Fig. 1:** Graphical representation for the text ‘The wolf chases the rabbit’, illustrating the need for edge labels.

A version of the graphical representation was produced to correspond to each coreference condition (Manual, Literal, CoreNLP, Illinois). For each story under each condition, the graphical representation contained an indexed list of unique entities detected. This was displayed as a unique number and the head noun of a
found entity. Coreferent mentions on the graph fragments were then annotated with the index of their corresponding entity, so the head noun could be read off of the list for question answering.

Two native English speakers were trained on how to read the graphical representations in order to answer the set of questions for each story under each coreference condition. More opinions were not necessary because answers to the set of questions were a matter of fact, does the graph contain an edge which represents an answer to the question and if so, what is that answer? A consensus on the answer to each of the questions was then reached by the two trained experts and the author, which can be justified by the principle of “what I tell you three times is true” [5, p.46].

This task was performed on graphical meaning representations, where the original text of the stories was not available to the two trained experts. The primary reason for this being that it appears readers of stories can mentally correct some coreference mistakes without realizing, in a similar manner to how people can still read a text which contains simple spelling mistakes [14].

A screen shot of this question answering interface for the manual coreference condition can be seen in Fig. 2. Referring mentions on the graph are appended with a dash and a single digit indicating the entity they refer to. These numbers and their corresponding entity are shown in the key in the top left corner. The blue rectangle with the text promiser on the link between the promise-01 and person_name_“Jupiter” -1 nodes shows an example tool-tip to describe the link type.

4 Intrinsic Coreference Evaluation

Table 1 shows the performance scores for the two coreference resolution systems on our dataset according to MUC, B³ and CEAFₑ, as well as the mean of the F1 scores as in [13]. The results of all three metrics suggest that CoreNLP outperforms the Illinois system on this dataset.

<table>
<thead>
<tr>
<th></th>
<th>MUC</th>
<th>B³</th>
<th>CEAFₑ</th>
<th>CoNLL F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CoreNLP</td>
<td>Illinois</td>
<td>CoreNLP</td>
<td>Illinois</td>
</tr>
<tr>
<td>Human</td>
<td>71.43</td>
<td>62.45</td>
<td>57.04</td>
<td>42.19</td>
</tr>
<tr>
<td>Animal</td>
<td>75.04</td>
<td>62.45</td>
<td>60.62</td>
<td>47.69</td>
</tr>
<tr>
<td>All</td>
<td>73.23</td>
<td>62.45</td>
<td>58.83</td>
<td>44.94</td>
</tr>
</tbody>
</table>

It is hard to give a broad overview of these results due to the variety of mistakes made by the systems. For example out of 20 sets of automatic coreference resolution results (two for each story) less than half contained the correct
Two native English speakers were trained on how to read the graphical representations in order to answer the set of questions for each story under each coreference condition. More opinions were not necessary because answers to the set of questions were a matter of fact, does the graph contain an edge which represents an answer to the question and if so, what is that answer? A consensus on the answer to each of the questions was then reached by the two trained experts and the author, which can be justified by the principle of "what I tell you three times is true" [5, p.46].

This task was performed on graphical meaning representations, where the original text of the stories was not available to the two trained experts. The primary reason for this being that it appears readers of stories can mentally correct some coreference mistakes without realizing, in a similar manner to how people can still read a text which contains simple spelling mistakes [14].

A screen shot of this question answering interface for the manual coreference condition can be seen in Fig. 2. Referring mentions on the graph are appended with a dash and a single digit indicating the entity they refer to. These numbers and their corresponding entity are shown in the key in the top left corner. The blue rectangle with the text "promiser" on the link between the "promise-01" and "person name "Jupiter" -1 nodes shows an example tool-tip to describe the link type.

4 Intrinsic Coreference Evaluation

Table 1 shows the performance scores for the two coreference resolution systems on our dataset according to MUC, B$_3$ and CEAFE, as well as the mean of the F1 scores as in [13]. The results of all three metrics suggest that CoreNLP outperforms the Illinois system on this dataset.

Table 1: Average F1 scores under commonly used evaluation metrics

<table>
<thead>
<tr>
<th></th>
<th>MUC</th>
<th>B$_3$</th>
<th>CEAFE</th>
<th>CoNLL F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>71.43</td>
<td>62.45</td>
<td>57.04</td>
<td>42.19</td>
</tr>
<tr>
<td>Animal</td>
<td>75.04</td>
<td>62.45</td>
<td>60.62</td>
<td>47.69</td>
</tr>
<tr>
<td>All</td>
<td>73.23</td>
<td>62.45</td>
<td>58.83</td>
<td>44.94</td>
</tr>
</tbody>
</table>

It is hard to give a broad overview of these results due to the variety of mistakes made by the systems. For example out of 20 sets of automatic coreference resolution results (two for each story) less than half contained the correct answer.
number of entities. In some cases, the correct number of entities simply meant that fewer correct entities were detected and additional spurious entities were introduced by the system. Furthermore, for some stories a single gold standard entity was detected as several entities in a system output. Other types of error relate to undetected or incorrectly resolved mentions.

Results are also given for the separate splits of human-only stories and stories with anthropomorphic animals. This is to show that using stories containing animals speaking as human characters had no detrimental affect on the performance of the coreference resolution systems on this dataset. Under the MUC and $B^3$ metrics the results in fact suggest that both systems performed at least as well on this split of the data. Table 1 shows that even on a small dataset, the differences in methods of calculation for these three metrics can noticeably affect the results. As such it is useful to compare the results under these metrics with the results of our task.

5 Extrinsic Coreference Evaluation

A rule for deciding if an answer should count as correct was introduced in order to be more lenient to the automatic coreference systems. There were multiple instances in which a mention required to answer a given question was not even detected by an automatic system but the node text for that mention, coupled with the output of the associated automatic coreference resolution system, would give the correct answer to that question. This is not necessarily the same as the answer that would be obtained from the meaning representation with no coreference resolution performed, as some other (correctly identified) entities may also be required to answer that question. A mention not being resolved should be distinguished from a mention being resolved incorrectly, which subsequently leads to the wrong answer being given for a question. For the results given here this leniency rule was applied more frequently to the Stanford system than the Illinois system, without which the performance gap between the two would be more pronounced. The following example question and corresponding answer under each coreference condition illustrates this rule and the effect of the different conditions on the ability to answer the questions. Consider:

**Question:** What did the father wish?

**Literal:** Not Applicable - as no pronouns are resolved, the question does not make sense in this instance.

**Manual:** For his sons to give his farm the same attention as he had given to the farm - ideal answer.

**CoreNLP:** For his sons to give his farm the same attention as he had given to it - where his and he correctly resolve to the father and it resolves to the same attention.

**Illinois:** For his sons to give the farm the same attention as he had given it - where his and he correctly resolve to the father and it is not resolved at all.
In this example the answer under the Literal condition is incorrect; as none of the pronouns are resolved it cannot be said that any of them refer to the father in the question. The answer using CoreNLP coreference information is counted as incorrect because it reads ‘For his sons to give his farm the same attention as he had given to the same attention’, which is nonsensical. The answer using coreference information from the Illinois system is counted as correct under the leniency rule previously described, as ‘it’ is neither correctly or incorrectly resolved but reads true to a human reader.

Table 2 gives the results of the question answering task under the four different coreference conditions. It is important to realize that given imperfect coreference resolution, a question may not make sense, and these cases are counted as incorrect answers. *Literal* refers to the results of answering the set of questions where no coreference resolution has been performed and the answer, where possible, is given simply by the text label(s) on the relevant node(s). This score represents the questions that can be answered without character information or concern parts of the text where characters are explicitly named. The *Manual* column represents the questions that can be answered with ideal, independently verified, coreference resolution. As the results show, all questions could be answered correctly using this version of the graphical representations, as if being answered from the original text. The *CoreNLP* and *Illinois* columns give the percentage of questions that could be correctly answered using the outputs from each automatic coreference resolution system respectively.

<table>
<thead>
<tr>
<th></th>
<th>Literal</th>
<th>Manual</th>
<th>CoreNLP</th>
<th>Illinois</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>27.27%</td>
<td>100%</td>
<td>36.36%</td>
<td>54.55%</td>
</tr>
<tr>
<td>Animal</td>
<td>35.00%</td>
<td>100%</td>
<td>50.00%</td>
<td>45.00%</td>
</tr>
<tr>
<td>Total</td>
<td>30.95%</td>
<td>100%</td>
<td>42.86%</td>
<td>50.00%</td>
</tr>
</tbody>
</table>

These extrinsic results show that nearly 43% of the questions could be answered when using Stanford CoreNLP and 50% when using Illinois’ system. Less than a third of the questions could be answered without any coreference resolution information, highlighting the importance of characters to understanding and reasoning about a text. The set of questions that can be answered with no coreference resolution performed is not a proper subset of those that can be correctly answered with either automatic coreference resolution system. The cause of this is that some explicitly named character mentions are incorrectly resolved to other characters by the automatic systems.

These results suggest that none of the intrinsic evaluation metrics discussed in this work are suitable for predicting a system’s usefulness to this task. Not only do the intrinsic metrics fail to accurately predict performance at this task, they give a misleading idea of which coreference resolution system would perform
better, suggesting that CoreNLP is the better of the two systems, while for this task the converse is true.

A contributing factor to the lack of correlation between the results of the question answering study and the coreference resolution evaluation metrics could be the incorrect detection of head nouns. Regardless of the precision and recall of the automatic systems under any of the three established metrics examined here, assigning an incorrect head noun to an entity can result in lowered performance for question answering. A very pronounced example of this occurs in one of the stories examined here, where the murderer character is assigned the head noun the man whom he murdered, leading to nonsensical answers to related questions. While only using head nouns may not give a true picture of the entities that they represent, they would in practice substitute co-referent mentions in order to answer questions and generating summaries. Steinberger et al. [16] performs such substitutions in order to correct dangling anaphoric expressions in generated summaries. For this dataset, although multiple entities were assigned non informative head words such as he by the automatic systems, it was never the case that a more specific co-referent mention in the same detected entity simply wasn’t assigned. Having an anaphor assigned as an entity’s head noun does little to aid the correct answering of questions related to this entity.

6 Conclusion

This study was designed to give the automatic coreference resolution systems the best possible score they could achieve on the described task. While it is not possible to define an exact level at which the extrinsic performance of these systems would be considered acceptable, the results of this task suggest they offer little improvement over not being used at all. Even then, they do not offer a clear cut improvement and introduce different types of error. Judging by the intrinsic performance of the systems, the extrinsic results were expected to be higher than the values we have reported. In theory, the results on our task should have been significantly higher if the systems’ primary loss of accuracy came from the incorrect resolution of minor characters. The questions used to evaluate these systems extrinsically concerned only the aspects of the stories deemed summary-worthy. As such, mistakes by the coreference resolution systems regarding minor characters are tolerable and would not affect the answers to the questions participants were asked. Conversely, coreference mistakes concerning major characters should have a very negative effect on their performance in this task. The results presented here then suggest that the areas where the automatic coreference resolution systems are failing concern the most critical characters.

The results discussed indicate that coreference information is needed, but automatic coreference resolution in its current state provides too large a source of error to be used for summarization and question answering based tasks, and that manual coreference resolution should be performed where feasible. The mistakes made by such systems appear to cover important aspects of a text, which
would seriously affect the automatic planning and generation of summaries. This work further suggests that the typically used established evaluation metrics for coreference resolution cannot predict a system’s usefulness to these tasks.

References


Hierarchical Clustering
of Estonian Verb Constructions

Dage Särg
Institute of Estonian and General Linguistics, University of Tartu, Estonia
dage.sarg@ut.ee

Abstract. The aim of this paper is to describe an experiment to cluster
Estonian multiword verb constructions in order to find out whether an
existing classification is related to the constructions’ sentence distributions. Constructions under consideration consist of two verbs, one finite
and one infinite. Agglomerative hierarchical clustering is performed on
the constructions, based on their sentence distributions. The morpholog-
ical composition of the clauses is used for features, and the results are
compared with a more frequently used lexicon-based approach.

1 Introduction

According to traditional descriptive grammar of Estonian [1], Estonian verb
constructions that consist of several verbs are divided into three types:

• compound tenses (see Fig. 1);
• chain verbs (periphrastic verbs that contain a finite + infinite verb, see Fig.
  2);
• verbs that have an infinite verb as a complement (Fig. 3).

While compound tenses are certainly a separate category, the distinction
between the two other types is not so clear: the descriptive Estonian Language
Grammar [1] lists the verbs and constructions that are considered as chain verbs,
and others that are considered to be verbs with an infinite verb as a complement.
This distinction is commonly taken as ground truth.

Chain verbs include modal verbs, as well as other more grammaticalised mul-
tiword verb constructions, while less grammaticalised (or non-grammaticalised)
multiword verb constructions remain a separate category. This means that gram-
maticalisation is viewed as a scalar phenomenon and constructions with verbs
on one end of the scale are considered chain verbs. The features defining an aux-
iliary or a verb in the middle of a grammaticalisation process, a semi-auxiliary,
have been described by Helle Metslang [2] as follows:

• it combines freely with any verb;
• it does not determine the presence or form of a subject in the clause;
• its only rection is the obligatory propositional complement;
• the construction that it forms does not combine freely with marked members of verb paradigm;
• it cannot appear as the head of the non-finite clause;
• it desemanticizes.

In this study, we attempt to find out if this distinction is also reflected in the sentence distributions of these two types of constructions. To do this, agglomerative hierarchical clustering of the clauses in which the constructions appear is used.

Although we use the term ‘multiword verb constructions’, it should be noted that this work concentrates on constructions that do not match the traditional definition of multiword expressions - e.g. as defined by Baldwin and Kim [4] who have stated that multiword expressions "have surprising properties not predicted by their component words". The meanings of constructions considered in this study are entirely predictable from the words that the expression consists of: the finite component of the construction gives the lexical meaning and the infinite component expresses modality, causativity, willingness or other properties.

The rest of the paper is organised as follows. In Section 2, the data and software that were used for the study are described. Section 3 gives an overview of the method: how the relevant dataset was extracted and processed, while in Section 4, the results are presented. After that, in Section 5, related research is described, and finally, Section 6 includes the discussion and explores the possibilities for future work on the topic.

Fig. 1. Example of a dependency tree containing a compound tense (in bold).

2 Data, Software

For this study, the Balanced Corpus of Estonian\(^1\) was used as the source of data. The corpus contains 15 million words of newspaper, fiction, and scientific texts.

\(^1\) http://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=en
In this study, we attempt to find out if this distinction is also reflected in the sentence distributions of these two types of constructions. To do this, agglomerative hierarchical clustering of the clauses in which the constructions appear is used.

Although we use the term 'multiword verb constructions', it should be noted that this work concentrates on constructions that do not match the traditional definition of multiword expressions - e.g. as defined by Baldwin and Kim [4] who have stated that multiword expressions "have surprising properties not predicted by their component words". The meanings of constructions considered in this study are entirely predictable from the words that the expression consists of: the finite component of the construction gives the lexical meaning and the infinite component expresses modality, causativity, willingness or other properties.

The rest of the paper is organised as follows. In Section 2, the data and software that were used for the study are described. Section 3 gives an overview of the method: how the relevant dataset was extracted and processed, while in Section 4, the results are presented. After that, in Section 5, related research is described, and finally, Section 6 includes the discussion and explores the possibilities for future work on the topic.

**Fig. 2.** Example of a dependency tree containing a chain verb (in bold). The chain verb consists of a finite verb form (‘hakkas’ as a phase verb meaning ‘started’) and of an infinite form (‘kasutama’ as a lexical verb meaning ‘use’).

"When out of fake tickets, Palm started to use Swedbank credit cards."

**Fig. 3.** Example of a dependency tree containing a multi-word verb construction (in bold) where the finite form (‘tahaksite’) is the predicate and the infinite form (‘unustada’) its complement (direct object).

"Which period in your life would you like to forget?"
(5 million words of each), and it is both morphologically and syntactically annotated. The morphological annotation has been done with Vabamorf\(^2\) tool which finds the correct analysis for more than 99% of tokens appearing in standard texts [5]. Syntactic parsing has been done with Estonian Constraint Grammar\(^3\) which performs dependency syntactic analysis with LAS of over 80% [6].

Multiword verb constructions in the corpus are annotated according to the distribution given by Estonian Language Grammar\([1]\) and can be seen from dependency trees on Figures 1–3. For compound tenses (Fig. 1) and chain verbs (Fig. 2), the infinite verb is the head (and the root of the whole clause) and the finite auxiliary verb is its dependent; for the last group (Fig. 3), the finite verb is annotated as the root and the infinite verb that serves as a complement is its dependent.

Clauses that contained the multiword verb constructions considered in this study were extracted from the dependency trees using the Python EstNLTK\(^4\) library, which is an NLTK adaption for Estonian, and vislcg3\(^5\), the parser that runs constraint grammar rulesets.

For clustering and visualisation, Scipy\(^6\) tools were used.

3 Method

3.1 Data Extraction

To use hierarchical clustering, the data needs to be presented in the form of feature vectors – each vector describes one construction and is essentially a set of feature frequencies characterising the construction. Our intention was to study sentence distribution of the constructions, but in compound and complex sentences we can intuitively assume that only word forms occurring in the same clause with the construction are relevant. Therefore, those sentences were split into clauses based on the clause boundaries that were annotated on the corpus with EstCG, and only the clauses containing the multiword verb construction were considered. In addition, the dependents of noun phrases were excluded from the clauses because they only modify the noun phrase and not the whole clause. Other words – both arguments and adjuncts – were all taken into account because there is no straightforward way for distinguishing them and we can assume that their relative importance in the clauses is reflected in the aggregated counts.

From the extracted clauses, the ones containing constructions that consisted of more than two verbs were discarded because most times, they cannot be classified straightforwardly into one construction type. The clauses where the verb construction did not consist of a finite and an infinite verb, were discarded as well because those were either serial verbs that are not in the scope of this study or mistakes coming from faulty clause boundaries.

\(^2\) [https://github.com/Filosoft/vabamorf](https://github.com/Filosoft/vabamorf)

\(^3\) [https://github.com/EstSyntax/EstCG](https://github.com/EstSyntax/EstCG)

\(^4\) [https://estnltk.github.io/estnltk/1.4.1/](https://estnltk.github.io/estnltk/1.4.1/)

\(^5\) [https://visl.sdu.dk/cg3.html](https://visl.sdu.dk/cg3.html)

\(^6\) [https://www.scipy.org/](https://www.scipy.org/)
3.2 Feature Selection and Reduction

Morphological analyses – combinations of the word’s part-of-speech tag and morphological features – were used for clustering. Although the corpus was syntactically annotated, the syntactic labels were not used as features. This is due to the fact that EstCG rules for syntactic annotation are uniformly based on the Estonian Language Grammar and therefore, the constructions are undoubtedly distinguished in the existing syntactic annotation the same way as in the Estonian Language Grammar.

Estonian is a morphologically rich language and the analysis given by the morphological analyser is very detailed; therefore the analysis was first simplified to reduce the number of different features. This included ignoring the following distinctive traits that were given out by the analyser but considered not relevant for this study:

- singular vs plural;
- 1st person vs 2nd person;
- proper nouns vs common nouns;
- positive vs comparative vs superlative adjectives;
- cardinal vs ordinal numbers;
- pronoun classes.

What remained in addition to part-of-speech-tags, was case information and distinction between 3rd person vs other (1st and 2nd person).

3.3 Composition of Feature Vectors

For each construction, the simplified morphological analyses of words appearing in the clauses, together with the construction, were counted. For example, from the sentence on Fig. 2, the features NOUN-ad (’lõppemisel’), NOUN-nom (’Palm’) and NOUN-part (’kreditteraat’) are counted as appearing once with the construction ‘hakkama’ + INF. The counts over all the clauses containing the constructions were aggregated and based on the counts, relative frequencies of analyses appearing together with the construction were calculated.

For further analysis, 38 most frequent constructions were chosen based on the count of clauses in which they appeared in the corpus. The most frequent construction appeared in more than 40 000 clauses in the corpus and the least frequent one under consideration appeared in 400 clauses. Constructions appearing in less than 400 clauses were not considered because the distribution of morphological analysis tags of constructions would probably be too arbitrary to use for clustering and drawing conclusions not regarding the specific clauses but their use in general.

3.4 Clustering

To the matrix of 38 constructions, where each row represents the relative frequencies of morphological tags appearing in clauses together with one construction,
agglomerative hierarchical clustering was performed using the Scipy clustering package. Cosine distance was used as a distance measure and complete linkage was the method used to calculate the distances between clusters.

4 Results

The clustering results are presented as a dendrogram on Figure 4. The dendrogram should be read from right to left: each construction is represented by a branch on the dendrogram and the point where two branches unite shows how closely related the two constructions are, based on the features described in the previous section.

It can be seen that constructions classified as chain verbs in the aforementioned descriptive grammar of Estonian [1] – underlined on the figure – do not form a uniform cluster but are rather scattered into different clusters. However, most of the clusters on the figure can still be interpreted, which means there is some signal in the chosen features.

The first cluster (light green, marked with number 1 on Fig. 4) at the top of the figure contains verbs of enabling and permission that require a patient expressed in the adessive case. The "palu"+inf (‘ask/beg’) construction, which is also connected to the cluster but more remotely, is also of this type. The next, "meeldi"+inf (‘like’) construction is farther from the others and this is the only construction on the figure that requires an agent expressed in the allative case.

The second, purplish cluster (number 2), is large and rather heterogeneous, however, the distances on the dendrogram between constructions are small. This cluster contains the most frequent modal constructions in Estonian ("pida"+sup (‘must/have to’), "vöi"+inf (‘can/may/might’), "saa"+inf (‘can/may’)), as well as some (almost) synonymous construction pairs, such as "suut"+inf (‘manage/be able’) and "joud"+inf (‘manage/be able’), or "püüd"+inf (‘try’) and "ürita"+inf (‘attempt’).

The third, light blue cluster (number 3) contains two chain verbs closely connected to each other, and four constructions related to willingness ("visitsi"+inf (‘bother’) and "taht"+inf (‘want’) or ableness ("oska"+inf (‘can’) and "tead"+inf (‘know’) to do something.

The fourth, red cluster is the hardest to interpret and seems most arbitrary of the clusters on the figure. The constructions "käi"+sup (‘go’) and "õnnestu"+inf (‘succeed’) are placed quite far from the others, as well as from each other which seems reasonable. "õnnestu"+inf (‘succeed’) and "tule"+inf (‘must/have to’) both can take an agent expressed in adessive case, but this is not obligatory. In fact, Pille Penjam [7] has stated that the "tule"+inf (‘must/have to’) construction has the adessive agent only in 20% of cases.

The fifth, green cluster includes two pairs of closely connected verb constructions: phase verb constructions "mine"+sup (‘go’) and "jää"+sup (‘stay’), and causative verb constructions "pane"+sup (‘make’) and "sundi"+sup (‘force’). In addition to all the numbered and coloured clusters containing several constructions, there is the last construction on the dendrogram - the "ole"+inf
Agglomerative hierarchical clustering was performed using the Scipy clustering package. Cosine distance was used as a distance measure and complete linkage was the method used to calculate the distances between clusters.

### Results

The clustering results are presented as a dendrogram on Figure 4. The dendrogram should be read from right to left: each construction is represented by a branch on the dendrogram and the point where two branches unite shows how closely related the two constructions are, based on the features described in the previous section.

It can be seen that constructions classified as chain verbs in the aforementioned descriptive grammar of Estonian [1] – underlined on the figure – do not form a uniform cluster but are rather scattered into different clusters. However, most of the clusters on the figure can still be interpreted, which means there is some signal in the chosen features.

The first cluster (light green, marked with number 1 on Fig. 4) at the top of the figure contains verbs of enabling and permission that require a patient expressed in the adessive case. The "palu"+inf ('ask/beg') construction, which is also connected to the cluster but more remotely, is also of this type. The next, "meeldi"+inf ('like') construction is farther from the others and this is the only construction on the figure that requires an agent expressed in the allative case.

The second, purplish cluster (number 2), is large and rather heterogeneous, however, the distances on the dendrogram between constructions are small. This cluster contains the most frequent modal constructions in Estonian ("pida"+sup ('must/have to'), "või"+inf ('can/may/might'), "saa"+inf ('can/may')), as well as some (almost) synonymous construction pairs, such as "suut"+inf ('manage/be able') and "jõud"+inf ('manage/be able') or "püüd"+inf ('try') and "õrit"+inf ('attempt').

The third, light blue cluster (number 3) contains two chain verbs closely connected to each other, and four constructions related to willingness ("vitsi"+inf ('bother') and "taht"+inf ('want') or ableness ("oska"+inf ('can') and "tead"+inf ('know')) to do something.

The fourth, red cluster is the hardest to interpret and seems most arbitrary of the clusters on the figure. The constructions "käi"+sup ('go') and "õnnestu"+inf ('succeed') are placed quite far from the others, as well as from each other which seems reasonable. "õnnestu"+inf ('succeed') and "tule"+inf ('must/have to') both can take an agent expressed in adessive case, but this is not obligatory. In fact, Pille Penjam [7] has stated that the "tule"+inf ('must/have to') construction has the adessive agent only in 20% of cases.

The fifth, green cluster includes two pairs of closely connected verb constructions: phase verb constructions "mine"+sup ('go') and "jää"+sup ('stay'), and causative verb constructions "pane"+sup ('make') and "sundi"+sup ('force').

In addition to all the numbered and coloured clusters containing several constructions, there is the last construction on the dendrogram - the "ole"+inf ('be') construction, which is the furthest cluster from all others on the figure. This construction is actually linguistically different from all the others, because in addition to multiword verbs, it also forms light verb constructions with adverbs or nouns which have not been removed from the dataset because their annotation in the treebank does not allow us to do this automatically.

Therefore, we can say that although the chain verbs do not form a cluster but belong to most clusters on the dendrogram, the clustering is not completely arbitrary.

---

**Fig. 4.** Clustered multi-word verb constructions. The constructions are noted with the lemma of the finite verb and 'inf' or 'sup' to signify whether the construction takes the infinitive or supine form of the infinite component. The supine/infinitive does not affect the forms of other words in the clause. The constructions considered to be chain verbs in the Estonian Language Grammar [1] are underlined.
5 Related Research

As stated in the introduction, the constructions viewed in this study are not multi-word expressions in the term’s traditional meaning. Therefore, in the literature overview, we are also not concentrating on works done on multiword expressions but on verbs and verb constructions.

5.1 Works on automatic verb classification

There are studies available on automatic verb classification and clustering for various languages, but, to our knowledge, none specifically about multiword verb constructions or clustering based on morphological analysis. Mostly, the goal of the classification is to divide the verbs in a language into syntactic classes (e.g. [8], [9]) or to achieve lexical verb classes similar to what Beth Levin ([10]) introduced for English (e.g. [11], [12], [13], [14], [15]).

For example, Merlo and Stevenson [8] have experimented with different supervised machine learning methods and hierarchical clustering to classify optionally intransitive English verbs into three classes that have different argument structures. For features, they used linguistical characteristics of the verbs, such as transitivity, causativity, animacy of a subject, etc, extracted automatically from a 29-million-word syntactically annotated corpus, using robust methods. They concluded that a small set of lexical features is sufficient to achieve a remarkable improvement in accuracy in comparison with the baseline.

Mayol et al. [9] have attempted a similar task for clustering Catalan verbs into syntactic classes - transitive, prepositional and intransitive - using a POS-tagged corpus of 16 million words. They chose the POS-tags of the words immediately following the verb for clustering, but also other linguistically motivated feature like whether the verb appears in passive constructions. For them, the method helped to successfully classify prototypical members of the classes.

Sabine Schulte im Walde has done extensive work on automatic acquisition of Levin-style semantic verb classes for German. In [11] and [12] she has explored the possibilities of using different features for verb class induction: syntactic frame types, prepositional phrase information, and selectional preferences. She found that using very fine-grained features did not improve the clustering results because in this case, the verbs’ individual differences started to have an influence.

Roberts and Egg [13] have experimented with selectional preference models to cluster German verbs: they used the classes of nouns that appear as verbs’ arguments for features. They found that all their selectional preference models were helpful in automatic verb clustering, but also the most simple approach using lexical preferences worked on their 880-million-word corpus only slightly worse than the others.

5.2 Works on Estonian verb constructions

There are several previous studies of Estonian verb constructions, but no attempts in automatic verb classification have been conducted. Heiki-Jaan Kaalep
and Kadri Muischneck [16] have written an overview of Estonian multiword expressions that contain a verb, and Kristel Uiboaeid [17] has done quantitative studies of verb constructions in Estonian dialects.

Siim Orasmaa [18] has studied verb subcategorization based on morphological information and found that the frequencies of morphological tags appearing in the same clause with a verb do give information about the subcategorization of the verb. However, he only considered clauses that contained one verb (finite main verb) to eliminate the possible arguments of other verbs.

6 Discussion and Future Work

Although the used method does not explicitly distinguish chain verbs from other multiword verb constructions, it still shows that the sentence distribution of verb constructions can, to some extent, be modelled based on morphological tags. This raises the question whether any other type of features would describe the situation better.

Roberts and Egg [13] mention that for German verb classification, a model based on nominal verb arguments works surprisingly well. Therefore, we decided to try clustering verb constructions based on the 5000 most frequent noun lemmas instead of morphological tags. The results were less interpretable than those of the main experiment: in the second experiment, the ”meeldi”+inf (‘like’) construction was still quite separated from all the others and two modal constructions appeared in the same cluster, but otherwise the clusters seemed arbitrary. This can be due to the corpus size: Robert’s and Egg’s corpus of 880 million words was almost 60 times larger than ours.

It could also be interesting to try clustering on generalised lexical features as described by Merlo and Stevenson [8] or on selectional preferences, but these possibilities remain out of scope for this study.

To improve the current clustering results, it would be possible to try other algorithms in addition to agglomerative hierarchical clustering. For example, Lin Sun and Anna Korhonen [19] have found that for clustering verbs into Levin-style classes, hierarchical graph factorization clustering performs better than agglomerative hierarchical clustering. In addition, it could be useful to try and combine the used features to reduce the dimensionality and sparsity of the frequency matrix: e.g. Schulte im Walde [12] has explained that it is crucial to find the appropriate level of feature specification - if the features are too specific, the verbs’ individual properties start to have too much of an influence.

References

2. Metslang, Helle. 2002. Iseseisev verb või abiverb? [Independent verb or auxiliary?] Presentation at a syntax workshop at the University of Tartu, Dec 6


In-domain or out-domain word embeddings?
A study for Legal Cases

Milagro Teruel and Cristian Cardellino
Universidad Nacional de Córdoba, Argentina
milagro.teruel@gmail.com ccardellino@famaf.unc.edu.ar

Abstract. In this paper we explore the contribution of word embeddings for domain adaptation, more precisely, for Named Entity Recognition and Classification in the legal domain. We compare two different kinds of models: obtained from a portion of Wikipedia, and obtained from a very small but in-domain corpus. Wikipedia models can be trained with a large corpus without further annotation efforts, but the examples used are out-of-domain. In contrast, in-domain models require new annotated examples and are expected to be more accurate, but also more prone to overfitting. In this setting, we expect that word embeddings will be useful because they provide a smoothed representation of the data. We compare different kinds of word embeddings with models trained with traditional linguistic features, and find that word embeddings decrease overall performance, but improve recall. This behavior is specially beneficial for legal applications, where coverage is more important than precision.

Keywords: domain adaptation, word embeddings, legal named entity recognition and classification

1 Introduction

Information Extraction (IE) systems can have a high impact on many tasks, and the legal domain is no exception [1]. Legal records are stored in natural language documents, unstructured or semi-structured, and constitute an important factor in law interpretation, legal reasoning and decision making. In particular, the identification of relevant jurisprudence allows law practitioners to build sounder argumentations for new cases, and many semi-automatic solutions are heavily used nowadays [2]. The task of Named Entity Recognition and Classification (NERC) is a building block on IE systems and can also influence the performance of other tasks related to the legal domain, such as argument mining, claim identification or automatic reasoning [3].

Although IE and NERC systems have been very popular over the last two decades and many systems are available, the processing of legal documents is special in several aspects. Legal documents are more structured than general text, often recurring to very formulaic expressions, the vocabulary is used with precise and special meaning, etc. Moreover, Named Entities are not only names of people, places or organizations, as in general-purpose NERC. Names of laws,
of typified procedures and even of concepts are also Named Entities in legal cases, as can be seen in the example in Figure 1.

Only very few annotated legal corpora exist, so it is difficult to train a Named Entity Recognizer. In the legal domain, the effort of annotation is specially high because only trained annotators can produce the corpora, given the very technical and precise semantics of those documents. A usual workaround consists in obtaining a model from general-domain documents, and then applying techniques of Domain Adaptation [4] improve the performance using the small available data in the target domain.

In this work, we propose to apply an automatic classifier for NERC in judgments of the European Court of Human Rights (ECHR). To cope with the lack of labeled examples, we use a portion of the English Wikipedia relevant to the legal domain as the starting training dataset. The selection of documents is obtained through the alignment of two ontologies: the legal ontology LKIF [5] which delimits the legal content of interest to us, and the YAGO ontology¹, which provides a way to link concepts to Wikipedia documents containing them. The use of a big number of Wikipedia documents allows us to have a baseline classifier for the legal domain without the need of an extensive annotation process. We were able to extract 102,000 entities and 4.5 million mentions.

For evaluation purposes, we have annotated a small set of cases of the ECHR. With this labeled dataset, we can compare the results of the classifier trained with a big number of Wikipedia documents against a classifier trained with a small amount in-domain data. Both classifiers, along with state of the art classifiers and simple heuristics, are applied to previously unseen judgments. Results are analyzed in Section 5, after the description of the dataset (Section 3) and the NERC system (Section 4).

2 Related work

A good review of data mining and information extraction methods applied to the legal domain can be found in the book by Stranieri (2005)[1]. It covers general Natural Language tools, as well as Machine Learning techniques.

Dozier et al. (2010) [2] approach the same problem as us, but with several differences. First, they consider fewer categories of Named Entities, leaving out abstractions and procedural entities. Second, their approach is rule-based, complemented with statistical methods, while we propose a fully data-driven method. Last, they use a different dataset: legal cases from United States Courts.

Quaresma and Gonçalves (2010) [6] also work in the NERC task over a set of European Union law documents in several languages, applying an automatic SVM classifier. They identify only locations, organizations, dates and references to other documents. Our proposal differs in the machine learning method and the use of out-of-domain training data.

Domain adaptation techniques have been also used for NERC in Social Media text by [7]. The authors gather unlabeled documents from several sources in

¹ www.yago-knowledge.org/
similar domains, together with a classifier pre-trained on a different domain. Next, they apply a bootstrapping method to select instances labeled with highest confidence by the classifier for further training.

The use of Wikipedia links in documents for Entity Identification is explored in [8], with successful results over in-domain and general documents.

3 Dataset descriptions

3.1 Wikipedia dataset

Wikipedia has been used as a corpus for NERC because it provides naturally occurring text where entities are manually tagged. Moreover, there is an explicit connection between Wikipedia URLs and nodes in the YAGO ontology. We aim to build a training dataset for NERC taking advantage of the vast amount of information already available in this resource. This process has two steps: the identification of entity mentions in documents, and the selection of relevant entities to include in the training dataset.

For the first task, we downloaded an XML dump of the English Wikipedia from March 2016. Then, we consider as mentions of an entity every anchor text of hyperlinks pointing to the entity’s Wikipedia page, as in [8]. This results in accurate examples, but with a high number of false negatives, provided that usually only the first mention of an entity in a document is labeled.

As we mentioned before, to select the relevant entities to train the model we rely on the legal ontology LKIF, which we aligned to the Wikipedia-linked ontology YAGO. From this ontology, we obtain 122 populated YAGO classes aligned to LKIF and all entities included in these classes. We extracted all articles that contained at least one link to an entity in this set, obtaining a total of 4.5 million mentions, corresponding to 102,000 unique entities. Then, we kept only sentences that contained at least one mention of a named entity.

For the NERC problem we consider each word as a training instance. Given a sentence, each word is labeled independently as a Named Entity if it is contained in the anchor of an entity mention. More than 90% of the words (instances) were not inside a mention. This imbalance in the classes results in largely biased classifiers, so we randomly subsampled non-named entity words to make them at most 50% of the corpus. The resulting corpus consists of 21 million words.

3.2 Abstraction levels

Once instances were obtained, we defined labels to assign them. There are multiple possible levels of abstraction for Named Entities. To assess the performance of a classifier in several abstraction levels, we established some orthogonal divisions in the LKIF-YAGO ontology, organized hierarchically. The final levels and the number of labels in each of them we use for classification are listed below:

1. NERC (6 labels): Instances are classified as: Abstraction, Act, Document, Organization, Person or O (Non-Entity).
NERC
The \( [\text{Court}_{\text{organization}}]_{\text{organization}} \) is not convinced by the reasoning of the \([\text{combined divisions of the Court of Cassation}]_{\text{organization}}\), because it was not indicated in the \([\text{judgment}]_{\text{abstraction}}\) that \([\text{E˘gitim-Sen}]_{\text{person}}\) had carried out \([\text{illegal activities}]_{\text{abstraction}}\)

LKIF
The \([\text{Court}]_{\text{Public_Body}}\) is not convinced by the reasoning of the \([\text{combined divisions of the Court of Cassation}]_{\text{Public_Body}}\), because it was not indicated in the \([\text{judgment}]_{\text{Decision}}\) that \([\text{E˘gitim-Sen}]_{\text{Legal_Person}}\) had carried out \([\text{illegal activities}]_{\text{Crime}}\)

YAGO
The \([\text{Court}]_{\text{wordnet_trial_court}}\) is not convinced by the reasoning of the \([\text{combined divisions of the Court of Cassation}]_{\text{wordnet_trial_court}}\) because it was not indicated in the \([\text{judgment}]_{\text{wordnet_judgment}}\) that \([\text{E˘gitim-Sen}]_{\text{wordnet_union}}\) had carried out \([\text{illegal activities}]_{\text{wordnet_illegality}}\)

Fig. 1. An example of legal entities annotated at different levels of granularity.

2. LKIF (21 labels): Instances are classified as belonging to an LKIF node, for example Legal_Role, Public_Body, Right, etc.
3. YAGO (122 labels): Instances are classified as belonging to the most concrete YAGO node possible, for example: wordnet_prosecutor,110484858, wordnet_human_right,105176846.

In Figure 1, we show an example annotated at these different levels of abstraction.

3.3 ECHR judgments dataset
From the ECHR website\(^2\) we downloaded 5 random documents and annotated the sections describing the relevant Law and the Court’s reasoning, leaving out the description of the facts. From a total of 19,000 words, we identified 1,500 entities and 3,650 instances.

The annotation was carried by four researchers involved in this work, following specific guidelines inspired in the LDC guidelines for annotation of NEs \([9]\). They were instructed to assign labels only at the YAGO level (the most specific), and the labels of the remaining levels were inferred using the hierarchical structure of the LKIF-YAGO ontology. Each annotator worked on at least two documents. To assess the agreement between annotators, four documents were annotated by two different people, achieving an inter-annotator from \(\kappa = .4\) to \(\kappa = .61\) using Cohen’s kappa coefficient \([10]\). Most of the disagreement between annotators was found for the recognition of concepts, not for their classification. We are working on developing the guidelines to enhance consistency among annotators. We will also apply automatic pre-processing and post-edition to an-

\(^2\) http://hudoc.echr.coe.int/eng
notated texts, in order to spot and correct errors. We randomly selected one of these duplicated annotations as correct for the final experimentation dataset.

To obtain high quality annotations independent of the current task, annotators were allowed to add new classes to the existing label list. Considering all final annotations, 14 new classes were added in the LKIF level and 73 on the YAGO level. This indicates the classes available in the LKIF-YAGO ontology are not covering the semantic domain of ECHR judgments, as can be expected in a domain adaptation problem. The main factor is that the LKIF ontology does not contain classes for the subdomain of Procedural Law or Crime, which are very frequent in a legal case document.

To assess performance in a realistic scenario, classifiers are evaluated in a single hold out document from the ECHR annotated dataset. The hold out document can contain previously unseen classes, as in a real application. All classifiers were trained using the remaining set of documents, separating 90% of instances for training and 10% for validation (parameter tuning). The evaluation process was repeated leaving out a different document in each iteration, and finally averaging the results of all iterations.

4 Building the NERC systems

4.1 Representation of instances

Word embeddings have been shown to help in cross-domain classification problems [11, 12] because they capture latent properties of words that are less dependent on the domain. This can also be viewed as a smoothing of the resulting representation, which should be specially adequate to address overfitting. It is also known that embeddings are more adequate the bigger the corpus they are learnt from, and if the corpus belongs to the same domain to which it will be applied. Therefore, we trained three kinds of embeddings: obtained from Wikipedia documents alone (a very big corpus), obtained from the judgments alone (an in-domain corpus), and obtained with a mixed corpus. The mixed corpus is composed of all the available judgments of the ECHR, and a similar quantity of text from Wikipedia (an augmented in-domain corpus).

To train the word embedding vectors we used the Word2Vec’s skip-gram algorithm [13]. For the Wikipedia dataset, we use the documents described in Section 3.1. Words with less than 5 occurrences were filtered out, resulting in a 2.5 million unique tokens, where the capitalization of words is preserved. To train word embeddings for judgments of the ECHR, we obtained all cases in English from the ECHR’s official site available on November 2016, summing up to 9,161 documents with 70 millions tokens and 131,000 unique words. The trained embeddings were vectors with 200 elements, and taking them we generated a matrix where each instance was represented by the vectors of the instance word and the vectors of the surrounding words by a symmetric window of 3 words at each size. If the word was near the beginning or the end of a sentence, or if any word was not in the Word2Vec model, the vector was padded with zeros.
We compared the performance of word embeddings with the standard set of features proposed by Finkel et al. [14] for the Stanford Parser CRF-model. For each instance we used: current word, current word PoS-tag, all the n-grams \((1 \leq n \leq 6)\) of characters forming the prefixes and suffixes of the word, the previous and next word, the bag of words (up to 4) at left and right, the tags of the surrounding sequence with a symmetric window of 2 words, and the presence of a word in as the total or as part of an entity in a gazetteer. To reduce the dimensionality of the final feature vector due to memory limitations, we applied a simple feature selection technique filtering out all features with variance less than \(2\times10^{-4}\). We call this representation handcrafted features, in contrast with automatically obtained word embeddings.

4.2 Classifiers

Using the corpus described in the previous section, we trained a Multilayer Perceptron (MLP) classifier for each abstraction level with a similar architecture. We experimented with one, two and three hidden layers, but it resulted that a single hidden layer performed better. We select an architecture with a hidden layer of size 8000 for the Wikipedia dataset with handcrafted features, of 5000 for the ECHR dataset with handcrafted features, and 2000 for all experiments using word embeddings.

To better assess the performance of these classifiers, we compare them with a sequential classifier: a Conditional Random Field model. We use the Stanford CRF for NERC [15] implementation. We retrained this classifier for all abstraction levels with the ECHR dataset, but the YAGO level had too many classes to be trained with the Wikipedia dataset. The representation used is the same as for MLP classifiers, except for the presence in gazetteers and the PoS tags of surrounding words. The second baseline proposed is a K-Nearest Neighbors classifier trained using the current, previous and following word tokens over the ECHR dataset. This is a very simple approach, equivalent to checking the overlap of the terms in the entity. We consider this baseline appropriate for the evaluation of the ECHR documents, where entity mentions tend to be more regular, such as "the applicant" or "the Court". However, the bigger vocabulary and higher number of entities made this approach unfeasible in the Wikipedia dataset.

5 Analysis of results

For this particular problem, accuracy does not throw much light upon the performance of the classifier because the performance for the majority class, \(non-NE\), eclipses the performance for the rest. To have a better insight on the performance, the metrics of precision and recall are more adequate. We calculated those metrics per class, and we provide a simple average not weighted by the population of the class (macro-average).

In total, we have trained four different MLP classifiers varying the representation of the instances used. The different representations are: handcrafted
Table 1. Performance of evaluation in ECHR holdout documents for the NERC task for classifiers trained with Wikipedia or ECHR documents. The metrics presented are averaged using a macro strategy. Classifiers are MultiLayer Perceptron (MLP), Conditional Random Field (CRF) and K-Nearest Neighbors (K-NN). The MLP classifier is also combined with Word Vectors (+WordEmb) representations from different datasets.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Wikipedia trained</th>
<th>ECHR trained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>MLP</td>
<td>.76</td>
<td>.56</td>
</tr>
<tr>
<td>MLP+WordEmb wiki</td>
<td>.73</td>
<td>.34</td>
</tr>
<tr>
<td>MLP+WordEmb mix</td>
<td>.75</td>
<td>.42</td>
</tr>
<tr>
<td>MLP+WordEmb echr</td>
<td>.75</td>
<td>.38</td>
</tr>
<tr>
<td>CRF</td>
<td>.73</td>
<td>.36</td>
</tr>
<tr>
<td>K-NN</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2. Performance of evaluation in ECHR holdout documents for the LKIF task for classifiers trained with Wikipedia or ECHR documents. The metrics presented are averaged using a macro strategy. Classifiers are MultiLayer Perceptron (MLP), Conditional Random Field (CRF) and K-Nearest Neighbors (K-NN). The MLP classifier is also combined with Word Vectors (+WordEmb) representations from different datasets.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Wikipedia trained</th>
<th>ECHR trained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>MLP</td>
<td>.76</td>
<td>.13</td>
</tr>
<tr>
<td>MLP+WordEmb wiki</td>
<td>.74</td>
<td>.08</td>
</tr>
<tr>
<td>MLP+WordEmb mix</td>
<td>.75</td>
<td>.10</td>
</tr>
<tr>
<td>MLP+WordEmb echr</td>
<td>.75</td>
<td>.11</td>
</tr>
<tr>
<td>CRF</td>
<td>.73</td>
<td>.07</td>
</tr>
<tr>
<td>K-NN</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3. Performance of evaluation in ECHR holdout documents for the YAGO task for classifiers trained with Wikipedia or ECHR documents. The metrics presented are averaged using a macro strategy. Classifiers are MultiLayer Perceptron (MLP), Conditional Random Field (CRF) and K-Nearest Neighbors (K-NN). The MLP classifier is also combined with Word Vectors (+WordEmb) representations from different datasets.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Wikipedia trained</th>
<th>ECHR trained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>MLP</td>
<td>.76</td>
<td>.06</td>
</tr>
<tr>
<td>MLP+WordEmb wiki</td>
<td>.74</td>
<td>.03</td>
</tr>
<tr>
<td>MLP+WordEmb mix</td>
<td>.75</td>
<td>.04</td>
</tr>
<tr>
<td>MLP+WordEmb echr</td>
<td>.74</td>
<td>.04</td>
</tr>
<tr>
<td>CRF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>K-NN</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
features, Wikipedia word embeddings, ECHR word embeddings, and word embeddings mixed from both sources. Additionally, we compare how the classifiers perform if they are trained with Wikipedia documents and ECHR documents. As a result, we evaluated eight MLP configurations. The performances of the baseline classifiers CRF and K-NN described in the previous section are also presented.

Results for the three proposed levels of abstraction: NERC, LKIF, and YAGO, are shown in Tables 1, 2, and 3 respectively. For the sake of comparison, we note that the reference tool for NERC achieves an F1 around 85-90% for in-domain training-testing with a large corpus at a level of granularity comparable to the NERC level [14]. In our approach performance never goes over 60% F1, probably due to a small training corpus, with few representatives of some of the classes. Indeed, using this same reference tool on our dataset yields only 56% performance at the NERC level.

We can observe from the three result tables that MLP classifiers outperform the K-NN baseline and perform comparably or better than the CRF classifier. Wikipedia-trained classifiers obtain lower figures in general, which could be expected because many of the classes in the ECHR corpus are not in Wikipedia and they could not be learned. This sets a ceiling to the performance of Wikipedia-trained models, which is a maximum accuracy of 0.866 for the LKIF level and 0.804 for the YAGO level. The obtained accuracy is then only 10 points below this ceiling of performance. However, even if the accuracy of Wikipedia-trained classifiers is not bad, the macro precision, recall and F1 score clearly show that they are not recognizing most of the classes. What they are actually doing is recognizing a small number of classes which have a big number of examples.

Focusing on ECHR-trained classifiers, we can see they achieve a better performance for all levels of abstraction, which is expected as we are training and evaluating on the same domain. However, developing a specific annotated corpus is costly, while Wikipedia provides a huge amount of annotated examples of a similar domain, for free.

Considering accuracy only, the MLP classifier performs better without word embeddings. As shown in figure 2, ECHR-trained classifiers with embeddings have a consistently higher recall, with a decrease in precision. This behavior is specially beneficial for legal applications, which are normally retrieval-related.

We also highlight that word embeddings trained with Wikipedia documents tend to perform better on models trained with the ECHR dataset, but there is no consistent difference between mixed and ECHR trained embeddings. The opposite occurs with the Wikipedia-trained models, where ECHR and mixed word embeddings improve both precision and recall. These two results show that, when we have a domain-specific model, embeddings obtained from a significantly bigger corpus are more beneficial. However, when no in-domain information is available, a representation obtained from many unlabeled examples improves more the classifiers. Even more, a very simple way of mixing examples for word embeddings in some cases enhances performance.
6 Discussion and future work

Our results show that word embeddings are beneficial to improve the recall of a small, in-domain model for NERC in legal documents. This is specially important for legal applications, which are mostly retrieval-centered and can tolerate noise better than silence. Word embeddings trained with unlabeled in-domain documents perform better than generic embeddings when the model has been trained in a different domain.

We have found that the most naïve combination of embeddings from different domains slightly improves classifier performance. We will investigate combinations of embeddings that are specifically targeted for domain adaptation.

At the same time, we have shown that Wikipedia-trained models achieve a reasonable level of performance in the legal domain, without any annotation cost. A promising line of work is to explore techniques to select documents of Wikipedia or other sources that will produce models closer to judgment documents, including more information of procedural law.

Acknowledgements

The authors have received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project MIREL: MIining and REasoning with Legal texts.

References

Discussion and future work

Our results show that word embeddings are beneficial to improve the recall of a small, in-domain model for NERC in legal documents. This is specially important for legal applications, which are mostly retrieval-centered and can tolerate noise better than silence. Word embeddings trained with unlabeled in-domain documents perform better than generic embeddings when the model has been trained in a different domain.

We have found that the most naïve combination of embeddings from different domains slightly improves classifier performance. We will investigate combinations of embeddings that are specifically targeted for domain adaptation. At the same time, we have shown that Wikipedia-trained models achieve a reasonable level of performance in the legal domain, without any annotation cost. A promising line of work is to explore techniques to select documents of Wikipedia or other sources that will produce models closer to judgment documents, including more information of procedural law.

Acknowledgements

The authors have received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project MIREL: Mining and REasoning with Legal texts.

References