Journal of Language Modelling

VOLUME 2 ISSUE 2
DECEMBER 2014

Institute of Computer Science
Polish Academy of Sciences
Warsaw

Journal of Language Modelling

VOLUME 2 ISSUE 2
DECEMBER 2014

JOURNAL OF
LANGUAGE MODELLING

ISSN 2299-8470 (electronic version)
ISSN 2299-856X (printed version)
http://jlm.ipipan.waw.pl/

Published by IPI PAN
Instytut Podstaw Informatyki
Polskiej Akademii Nauk
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland

Circulation: 100 + print on demand

Layout designed by Adam Twardoch.
Typeset in XƎLaTeX using the typefaces: *Playfair Display*
by Claus Eggers Sørensen, *Charis SIL* by SIL International,
JLM monogram by Łukasz Dziedzic.

A cognitively plausible model for grammar induction

Roni Katzir
Tel Aviv University

ABSTRACT

This paper aims to bring theoretical linguistics and cognition-general theories of learning into closer contact. I argue that linguists' notions of rich Universal Grammars (UGs) are well-founded, but that cognition-general learning approaches are viable as well and that the two can and should co-exist and support each other. Specifically, I use the observation that any theory of UG provides a learning criterion – the total memory space used to store a grammar and its encoding of the input – that supports learning according to the principle of Minimum Description-Length. This mapping from UGs to learners maintains a minimal ontological commitment: the learner for a particular UG uses only what is already required to account for linguistic competence in adults. I suggest that such learners should be our null hypothesis regarding the child's learning mechanism, and that furthermore, the mapping from theories of UG to learners provides a framework for comparing theories of UG.

Keywords: Universal Grammar, learning, Minimum Description-Length

1 INTRODUCTION

A central task in theoretical linguistics (TL) is constructing theories of competence – grammars (alternatively seen as computer programs) that have an opinion (a simple yes/no or a more fine-grained evaluation) about possible inputs. A broader goal of TL is characterizing the range of possible grammars that adult speakers can have. Thus, linguists agree that humans can mentally represent grammars from a set of possible candidates and use these grammars to analyze inputs.

Of course, much disagreement remains about the correct competence theories and the characterization of the range of theories. The characterization of the range of allowable grammars – which can be thought of as a reference machine into which individual grammars are written – is often referred to as Universal Grammar (UG).[1] Starting with UG, the child reaches a particular grammar through exposure to a linguistic environment. As pointed out by Chomsky (1965), this view assigns a central role to learnability in investigating UG: a linguistic theory must specify a range of grammars that can be attained using the cognitive machinery and data available to the child. Moreover, UG can provide an evaluation metric that allows the child to compare potential grammars given the data. In its original formulation, this evaluation metric was stated in terms of simplicity, a notion that – though defined with respect to a concrete UG – is also often seen as a cognition-general (CG) principle.

One might hope, then, that TL theories of competence and CG theories of learning would have a close relationship: that theories of UG would map onto theories of learning through an evaluation metric, and that theories of learning would restrict the choice of UG. In practice, however, the evaluation metric has been largely abandoned, and the two domains have never succeeded in constraining one another. Worse, TL and CG approaches have grown to be considered mutually incompatible. There are various different aspects to this ostensible incompatibility, such as whether linguistic knowledge involves structured, rule-like representations or not, whether probabilities play a role, and so on. Perhaps most fundamental among the perceived differences is how the two approaches view learning. TL, following a more hopeful beginning, has adopted a deeply skeptical stance that rejects the possibility of any meaningful learning and relegates most of the linguistic ability of adults to the innate component, and often to UG itself (that is, to the reference machine). CG, on the other hand,

[1] Elsewhere in the literature, UG is sometimes used to refer to the range of possible grammars (rather than to its intentional characterization as a reference machine), and sometimes it is used to refer to the combination of the range of possible grammars and the learning mechanism. Here UG will refer strictly to the reference machine. The term UG has sometimes been associated with approaches that assume a substantial innate component. Here I will use it neutrally – this paper makes no claims as to the correct theory of UG.

tends to be confident of learning and skeptical of the innate component (and especially of UG). The perceived incompatibility between TL and CG has led over the years to a growing divide between the two disciplines.

Over the past decade or so, the Bayesian program for cognition and the closely related framework of Minimum Description-Length (MDL) have brought the two disciplines closer by articulating CG views that can integrate probabilistic reasoning with structured, symbolic representations. In the other direction, proposals such as Marcus (2000) and Yang (2004, 2010) offer TL perspectives that connect with CG approaches to learning. But a sizable gap remains: even CG-oriented TL proposals such as those of Marcus and Yang still question the ability of general learning mechanisms to generalize correctly from the data, embracing instead restrictive theories of the innate component; and even TL-oriented CG proposals such as Goldsmith (2001), Dowman (2007), Foraker *et al.* (2009), and Perfors *et al.* (2011) still emphasize the power of general-purpose learning mechanisms and question whether the innate component should be quite as rich as TL would have it.

This paper has two goals. First, I wish to explain why the skepticism in both directions is misguided. In particular, I will explain why linguists believe in a complex innate component – including a nontrivial UG – even in the face of powerful statistical learners. I will do this by presenting two kinds of evidence that linguists rely on that have nothing to do with questions of learnability in principle. I will also explain why many cognitive scientists are confident that learning is a real possibility, despite the arguments against learning in the TL literature. My second goal is to offer a TL view that treats the learnable and the innate as mutually supportive rather than conflicting. The ability of CG mechanisms to learn, on this view, is interpreted not as a reason to reduce the innate component – though it will be a reason to bring back into consideration theories that leave much to be learned – but rather as a way to extract nuanced predictions from competing theories of that component.

I start, in Section 2, by reviewing the history of the divide between TL and CG, focusing first on the roots of TL pessimism regarding learning (Section 2.1) and then on CG optimism regarding the same (Section 2.2). In Section 3 I evaluate the two positions and argue that TL was wrong to dismiss learning but right to emphasize potentially restrictive UGs, while CG was right to emphasize learning but wrong to dismiss potentially restrictive UGs (Section 3.1). In Section 3.2 I explain how the co-existence of rich UGs and meaningful learning is not only possible but in fact a good state of affairs, one that allows us to revive the old hope of mutual collaboration from the early days of generative grammar. In Section 3.3 I explain how any fully explicit theory of UG provides us with a CG learner – specifically, a Minimum Description-Length (MDL) learner – and that this provides both a starting point for the study of learning and a basis for comparing competing theories of UG. Section 4 illustrates this mapping from UG to MDL learner using a simple UG and a couple of toy examples. Section 5 concludes.

2 TL AND CG: A BRIEF HISTORY OF THE SCHISM

2.1 *TL: Skepticism about learning*

2.1.1 Identification in the limit

In an influential paper, Gold (1967) introduced a learning paradigm, *identification in the limit (iitl)*, and proved that learning of this kind is impossible even in seemingly simple cases. In *iitl*, a learner g is presented with a sequence (or *text*) T of elements from a language L, where L is known to be taken from a set C of candidate languages. After each new element in T is presented, g guesses a language in C. If after a certain point all of g's guesses are the same correct guess (in this case, L), we will say that g has identified L in the limit from T. If g can identify in the limit any $L \in C$ based on any *fair* text in L (that is, a text in L in which every $w \in L$ appears at some point, and in which nothing appears that is not in L), we will say that g identifies C in the limit. If such a g exists, we will say that C is identifiable in the limit.

Certain simple families of languages are *iitl*. For example, the set of all finite languages over a finite alphabet Σ is *iitl*: if g guesses at each point the language that is the union of all the elements in T that have been encountered so far, it will always identify the source language in the limit. Similarly, any C that can be written as $\{L_i | i \geq 1\}$, where $L_i \subset L_{i+1}$ for all i, is *iitl*: g can identify C in the limit by always guessing the minimal L_i that contains all the elements in T that have been encountered so far. Changing these families of languages only slightly makes them not *iitl*. For example, adding a single infinite language to the set of all finite languages makes the set not *iitl*. In the second, more general example, adding $L_\infty = \bigcup L_i$ to C makes the result (as well as any set that contains it) not *iitl*. To see why, assume to the contrary that $C' = C \cup \{L_\infty\}$ is *iitl*. Let g be a learner that identifies C' in the limit. We can construct a text T that starts as a text in L_1 up until the first point where g guesses L_1 (such a point exists by assumption), continues as a text in L_2 up until the first following point where g guesses L_2, then continues as a text in L_3 until g guesses L_3, and so on. The result is a text in L_∞, but g makes infinitely many different guesses and so never converges on a correct answer, contrary to assumption.

Gold's setting rules out learning even in intuitively very simple families of languages, like the set of all regular languages.[2] For theoretical linguists, this has confirmed a growing skepticism (already discussed explicitly in Chomsky 1965, pp. 56–58) about the role of learning in linguistic competence. The skepticism was grounded in a general sense that learning is hard and that the data available to the child are insufficient. Gold's results can be seen as providing formal justification for this skepticism: assuming *iitl* is an appropriate model for language learning in humans, the set of possible languages must be severely restricted. Osherson *et al.* (1984) formulate further assumptions about human learning that, if correct, would entail an even more restrictive version UG in which the task of the learner is reduced to choosing from a finite set of candidate languages. Examples of linguistic approaches that adopt the finite version of UG are the Principles and Parameters framework of Generative Grammar (P&P; Chomsky 1981) and Optimality Theory (OT; Prince and Smolensky 1993).

It is worth noting that, while a restricted enough UG addresses the theoretical problem of *iitl*, even the finite version does not guaran-

[2] A full characterization of when a family of languages is *iitl* is provided by Angluin (1980). Algorithms that guarantee *iitl* for various classes of languages include Angluin (1982), Koshiba *et al.* (1997), Clark and Eyraud (2007), Heinz (2010), and Yoshinaka (2011). Note that arguments such as Gold's show that, under the relevant assumptions, *no* learner can succeed. This is a stronger result than showing that a particular learner cannot succeed (such as the problem identified by Braine 1971, Baker 1979, and Dell 1981 for the specific evaluation metric of Chomsky and Halle 1968).

tee an easy task in practice, since a finite space can still be dauntingly large. In the P&P framework, for example, there are 2^n settings, where n is the number of parameters (on the standard assumption that parameters are binary), and in OT there are $n!$ different constraint rankings, where n is the number of constraints. Noise and cognitive limitations further complicate the task. See Clark and Roberts (1993), Gibson and Wexler (1994), Niyogi and Berwick (1996), and Yang (2002) for attempts to tackle the practical issues of acquisition within P&P and Tesar and Smolensky (1998), Boersma and Hayes (2001), and Magri (2013) for a similar discussion within OT.

2.1.2 Poverty of the stimulus

Much of the disagreement between TL and CG has centered on a form of argument known as the argument from the poverty of the stimulus (POS), involving some property P that humans demonstrate in their language in spite of apparently insufficient support for P in the data. To cite a well-known (and highly controversial) example, English-speaking children will form a yes/no question by fronting the structurally highest auxiliary rather than the leftmost one, thus forming the yes/no interrogative version of *The monkey that is jumping can sing* by asking *Can the monkey that is jumping sing?* rather than **Is the monkey that jumping can sing?* (where $*$ marks ungrammaticality). They do so, it appears, despite hearing only simpler yes/no questions such as *Is the monkey jumping?* (from *The monkey is jumping*) and *Can the monkey sing?* (from *The monkey can sing*), where structurally highest and leftmost amount to the same thing. This has been taken to show that the innate component ensures this choice by making available structure-dependent generalizations but not rules that depend on linear order. See Berwick *et al.* (2011) and Clark and Lappin (2011), as well as references therein, for discussion.

While the form of POS arguments is clear enough, it is often difficult to establish any particular POS argument for humans in practice, even in a simple case such as the one just mentioned.[3] For example, how can we determine just what kind of evidence would suffice to make the relevant choice empirically? Could there be indirect

[3] In organisms for which it is possible to conduct controlled POS experiments, the situation is different, as Dyer and Dickinson (1994)'s work on honeybees shows.

sources of information that would predispose the child against forming ordering-based generalizations? And how sure are we that we know exactly what data the subjects have encountered over those few years prior to the experiment? Some progress has been made on these questions (see Legate and Yang (2002), Lidz *et al.* (2003), Yang (2010), and Hsu and Chater (2010) for thoughts on quantifying the information available to the the child; see Crain and Pietroski (2002) for how POS can be constructed from developmental stages in which children exhibit very specific linguistic knowledge that is incompatible with their ambient language but compatible with other natural languages; and see Wilson (2006) for an experimental paradigm designed to test the child's generalization beyond the data in POS situations), but the core weakness of relying on what we think can be learned and what we think the child hears – two questions that can be prohibitively difficult to answer – remains.[4]

2.1.3 Richness of the stimulus

If children can be shown to systematically *not* demonstrate a property P in their language despite an adequate amount of evidence supporting P in the input, we can conclude that this failure is due to the innate component. We can term such evidence an argument from the richness of the stimulus (ROS).[5] For example, Peña *et al.* (2002) have shown that, while humans are capable of extracting abstract dependencies within words, they fail on this task when combined with a segmentation task (a task that subjects perform well on, both on its own and when combined with the task of extracting word-internal dependencies). Similarly, Moreton (2008) has shown that humans are significantly better at learning certain phonological dependencies – specifically, dependencies relating the height of the vowels in two adjacent syllables – than other phonological dependencies – dependencies relating the height of a vowel to the voicedness of the following consonant

[4] This is not to say that the POS argument above has been shown to be incorrect. Despite multiple attempts to do so in the CG literature, the POS argument using subject-auxiliary inversion remains an open question. See Berwick *et al.* (2011) for relevant discussion.

[5] See Smith (1966) for an early example of this kind of argument in humans, and see Garcia *et al.* (1974) for a particularly clear example of the argument in rats.

and dependencies relating the voicedness of consonants in two adjacent syllables – even though the two patterns are equally prominent perceptually and are both abundantly represented in the input.

One must ensure, of course, that prior exposure has not biased the subjects against observing the relevant patterns. This, however, is considerably easier in practice than the reverse task, essential to POS, of ensuring that a certain pattern is never attested in the data. And as the above examples show – see Bonatti *et al.* (2005), Endress *et al.* (2007), Endress and Mehler (2010), Becker *et al.* (2011), and Hunter and Lidz (2013), among others, for further evidence of this kind – ROS lends itself to the design of controlled experiments that can inform us about what humans fail to learn.

2.1.4 Typology

Perhaps the most common source for enrichments of the innate component comes from the routine TL task of examining individual languages and comparing the results across a range of languages. If language after language shows the same property P (which can be an absolute universal, such as "Has nouns" or an implicational universal, such as "If demonstratives and adjectives precede the noun, then demonstratives precede adjectives"), we can sometimes conclude that P is due to the innate component.

As usual, caution is needed: for some properties, other sources, such as communication pressure, might be responsible rather than the innate component. For example, $P =$ "Verbs have a small number of arguments" or $P =$ "Has vowels". More interestingly, P may arise not through any direct benefit to the speakers but as properties that enhance the transmission of language between generations of speakers. See Kirby (2000, 2002); Kirby *et al.* (2004); Smith *et al.* (2003) as well as Niyogi and Berwick (1997, 2009). Less frequently, P can be explained away by appealing to historical accident.[6]

[6] Controlling completely for historical accident is quite challenging in practice, but the emergence of the Nicaraguan Sign Language (Senghas *et al.* 2004) and of the Al-Sayyid Bedouin Sign Language (Sandler *et al.* 2005) provide an approximation. In non-human species it is sometimes possible to explore typological questions in lab settings that control in full for historical accident, as shown by the work of Feher *et al.* (2009) on the emergence of typical song patterns in zebra finches over several generations, starting from birds grown in isolation.

But in many cases, *P* has little if anything to recommend it in terms of communication efficiency and other functionalist criteria. Suppose, to take a syntactic example discovered by Ross (1967), that I heard you say that you saw Max and some lady at the party last night, but I don't know the identity of the lady in question. I could use a roundabout inquiry such as *I heard you saw Max and some lady at the party; can you tell me which lady?*, or I could use a paraphrase such as *Which lady did you see Max with _ at the party?*, where the conjunction *and* in the original sentence is replaced with the preposition *with*. But what I cannot do, in English or in any other known language, is use the standard way to form a question and say **Which lady did you see Max and _ at the party last night?*, despite its obvious usefulness for the conversation (*P* in this case could be "Does not allow a question to target a single conjunct"). To cite a different example, discovered by Horn (1972), no natural language has a connective corresponding to *NAND* (= not and) or a quantificational determiner corresponding to *NALL* (= not all), despite the usefulness of these concepts in everyday life (as well as in artificial settings).[7] In such cases, it seems reasonable to ensure *P* through the innate component.[8, 9]

2.2 *CG: Optimism about learning*

2.2.1 The probabilistic turn

Other work, both theoretical and experimental, supports a less restrictive view on learning than the TL view. First, as has often been ob-

[7] See Horn (2011) and Katzir and Singh (2013) for discussion of the general context of this typological fact.

[8] Evans and Levinson (2009) and Levinson and Evans (2010) have made the remarkable claim that language universals do not exist. They do not discuss the Ross (1967)'s and Horn (1972)'s cases discussed above. See the commentaries following Evans and Levinson (2009), as well as Abels and Neeleman (2010), Crain *et al.* (2010), Reuland and Everaert (2010), Harbour (2011), and Matthewson (2012), among others, for additional problems with Evans and Levinson's claim.

[9] The discussion in this subsection is framed as one about absolute properties. See Tily and Jaeger (2011) and Piantadosi and Gibson (2013) for discussion of the challenges of obtaining a large enough sample to establish such universals statistically. In addition to absolute universals, quantitative typological evidence offers a rich source of information for TL, though using this information is still difficult at present. See Sauerland and Bobaljik (2013) for an interesting example.

served, some of Gold's assumptions do not seem to match the situation of the human language learner. In particular, the learner in *iitl* is expected to guess perfectly based on any fair text in the target language. No provision is made for discounting (or excluding completely) texts that are in some sense deviant, and no guess that is less than perfect counts. In acquisition, on the other hand, it is far from obvious that all sequences of inputs are equally good, and learning may well count as successful even if the child ends up having somewhat different judgments from its parents' about various sentences.[10] Relaxing this requirement, as has been done in the probabilistic settings of Horning (1969) and others, yields notions of learning that are often much more inclusive than *iitl*. Horning's setting involves the same form of text presentation as Gold's, but the texts are generated by taking independent, identically distributed samples from the strings generated by a probabilistic context-free grammar (PCFG), and the criterion for learning is modified. On these assumptions, the set of languages generated by PCFGs is learnable, even though the set of languages generated by Context-Free Grammars (CFGs) is not *iitl*.

Horning's results – and those of later probabilistic developments such as Wexler and Culicover (1980), Osherson *et al.* (1986), Angluin (1988), Kapur (1991), and Chater and Vitányi (2007) – can be seen as evidence that a probabilistic approach is both more natural and more successful than *iitl*.[11] Experimental data about specific learning

[10] A different aspect of *iitl* that could be changed with significant consequences for learnability is the assumption that the learner is only exposed to positive evidence. If the learner is exposed both to positive and to negative evidence (for example, as a sequence of strings paired with a grammaticality judgment), many more families of languages become learnable, including families that might be of potential linguistic interest. (Intuitively, the reason negative evidence helps is that it breaks all the subset relations between the languages in C – see Gold 1967 for discussion.) Unfortunately, infants do not seem to have access to anything like systematic negative evidence (Brown and Hanlon 1970; Marcus 1993).

[11] Care must be taken, however, in interpreting positive results about such models from the perspective of language acquisition. Horning (1969)'s original result applies to (unambiguous) PCFGs, a class of grammars that is not a realistic model of natural languages. Osherson *et al.* (1986) prove that a much broader class of languages can be identified with probability one from a similar form of text presentation (that is, through independent identically distributed draws from the language; see Clark 2001 for further extension). However, this result

tasks has provided empirical evidence for the role of statistics in learning, as well as further clarification of the requirements for a successful theory of learning in humans. One example is the segmentation experiments of Saffran *et al.* (1996), who showed that infants can reliably segment an artificially-generated input after a short exposure.[12] Since the only cues for segmentation in these experiments are statistical, we can conclude that a learner must be able to make use of statistical regularities in the input. In addition, these results show that a model for human learning should succeed even with unsegmented input.[13] Finally, the success of the babies in learning after such a brief exposure provides a preliminary quantitative measure of the performance of the learner. Further evidence that humans are skillful statistical learners come from Sobel *et al.* (2004) and Griffiths and Tenenbaum (2006), among others, who demonstrate the sensitivity of humans (both children and adults) to statistical information.

2.2.2 Task-specific approaches

Experimental results about learning tasks, of the kind mentioned above, have sometimes inspired task-specific (but domain-general) learning models: relatively simple mechanisms, usually sensitive to statistics, that form part of a CG toolkit. For example, the results of Saffran *et al.*, as well as those of subsequent experiments within the paradigm, have been taken to show that humans can employ certain segmentation techniques. One mechanism, based on Harris (1955) and suggested as the mechanism behind the infant segmentation data by Aslin *et al.* (1998), involves the tracking of transitional probabilities

requires knowing the possible distributions. If this assumption is replaced by more realistic requirements, the classes of languages that can be identified become considerably more limited, as shown by Angluin (1988) and Pitt (1989). In fact, if the child is required to perform distribution-free learning with probability one, the classes of languages that are identifiable revert to those that are Gold-identifiable. See Niyogi (2006) and Clark and Lappin (2011) for further discussion.

[12] Other examples include the tasks of categorization, the learning of phonotactics, and the induction of grammatical rules.

[13] Removing the segmentation marks in the text makes the learning problem harder. For example, the family $C = \{\{a\}, \{aa\}\}$ is trivial to learn from a segmented text but impossible to learn from an unsegmented text. Both Gold and Horning require the input to be segmented.

between syllables. Transitions tend to be more restrictive within words than across words, so segmentation can proceed by finding drops in transitional probability. Different task-specific models of segmentation have been offered by Brent and Cartwright (1996), Christiansen *et al.* (1998), Brent (1999), Mattys *et al.* (1999), Johnson and Jusczyk (2001), Venkataraman (2001), and Batchelder (2002), among others. Other task-specific (but potentially domain-general) learning mechanisms that have been proposed in the literature include mechanisms for processing identity relations (Endress *et al.* 2007) and positional relations (Endress and Mehler 2009).[14]

2.2.3 Prediction and description length

Another CG approach, one that is radically different from the task-specific approach – and the one I will try to support in this paper – is the idea of learning everything at once, with particular learning tasks (such as segmentation, categorization, syntactic learning, and so on) arising as by-products of a very general learning process. Here a principled approach is provided by the theory of prediction developed by Solomonoff (1964).[15] Simplifying, we consider all the different hypotheses about the data, each treated as a computer program that outputs the data, and we evaluate each hypothesis according to its length. The learner bases its guesses about the continuation of the input based on a weighted sum of all the hypotheses compatible with the observations so far, with shorter hypotheses receiving higher weights. Recently, this approach has been proposed by Chater and Vitányi (2007) and Hsu *et al.* (2011) as a useful abstraction – a form of *ideal learning* – for evaluating certain claims about the learnability of natural language.

While fully general and mathematically sound, ideal learning as originally formalized is not cognitively plausible, nor is it meant to be. In its pure form, ideal learning is not even computable (though see Solomonoff 2008 for thoughts on how to address this concern).

[14] See Endress *et al.* (2009) and Endress and Bonatti (2013) for further discussion of such mechanisms and qualifications of their generality.

[15] Related notions were developed by Kolmogorov (1965) and Chaitin (1966). See Li and Vitányi (1997) for discussion. Learning of this kind is guaranteed to minimize errors in a certain sense, as shown by Solomonoff (1978) and Chater and Vitányi (2007).

Another challenge to making Chater and Vitányi's model cognitively plausible is that it is stated with respect to a very broad UG – in its original form, a Turing-complete UG (which is the source of the non-computability). If we wish to take into account arguments for a more restrictive innate component, such as the arguments from ROS and from the typology, we should re-state Chater and Vitányi's model in terms of more limited UGs. Restricting the set of hypotheses can both ensure computability and make the model work with linguistically realistic UGs, but the computations required to derive the predictions in a Solomonoff-based ideal learner such as Chater and Vitányi's can still be prohibitively complex.

The approximation to Kolmogorov Complexity known as Minimum Description-Length (MDL; Rissanen 1978) offers a way to overcome the difficulties of ideal learning while maintaining both the weighting of hypotheses according to their length and the idea of general learning, with particular tasks falling out as by-products.[16] In MDL – and in the closely related Bayesian framework – the hypothesis space is restricted, and the search aims at finding a single hypothesis that minimizes the total description length (or, in the Bayesian framework, a hypothesis that maximizes the posterior probability). MDL has been used for grammar induction in the works of Berwick (1982), Rissanen and Ristad (1994), Stolcke (1994), Brent and Cartwright (1996), Chen (1996), Grünwald (1996), de Marcken (1996), Osborne and Briscoe (1997), Brent (1999), Clark (2001), Goldsmith (2001), Onnis *et al.* (2002), Zuidema (2003), Dowman (2007), Chang (2008), and Rasin and Katzir (2013) among others. In Section 3.3 I will suggest that MDL arises as a natural criterion for the evaluation of grammars given the data – and thus as a natural CG learning mechanism – from the commitment to an explicit UG made in TL.

[16] See also the closely related approach known as Minimum Message Length (MML; Wallace and Boulton 1968). An approach related to MDL and MML is the search for a grammar (usually a context-free grammar) that generates the input data as its only possible output. The problem of finding such a grammar – the so-called *shortest grammar problem* – has its roots in Lempel and Ziv (1976) and has been studied by Nevill-Manning and Witten (1997), Kieffer and Yang (2000), Charikar *et al.* (2005), and Dębowski (2011), among others.

3 REASSESSMENT

3.1 *A rich UG and the possibility of learning both exist*

As we saw, TL has good reasons to assume a nontrivial UG: while *iitl* seems inapplicable to the condition of the child, and while POS arguments are susceptible to successful learning models, ROS and typological arguments do not depend on learnability in principle. Indeed, the better the general-purpose mechanisms that one can assume, the more surprising both failures to learn and systematic typological patterns become. At the same time, the CG models of learning are clearly very much an option. None of the arguments against learning in principle holds, and it seems that humans are quite good at learning statistical distributions (as shown by Sobel *et al.* 2004 and Griffiths and Tenenbaum 2006, among others).

Assuming that (almost) everything is innate or that (almost) everything is learned was perhaps convenient at one point as a working hypothesis: if we already have an elaborate innate component, we might hope that we could do without a sophisticated learning mechanism, and vice versa. But a rich innate component and a powerful CG mechanism are not logically incompatible, and it is worth noting that the state of the art in each project still leaves a significant amount of work for the other. At the very least, then, the two respective research projects should continue to co-exist: TL should keep studying the innate component focusing on ROS and typological evidence, perhaps showing more caution with POS arguments than it did before; and CG should keep studying what humans can learn and how, perhaps showing a better appreciation for the role of innateness in shaping adult linguistic abilities.

But there is also a more interesting option, one that allows a tighter collaboration between the two research projects and that enables discoveries in one to translate into tools for the other. This option, a hope from the early days of generative grammar, was made possible by the advent of the Bayesian program for cognition and of the closely related MDL framework, both of which allow the integration of structured representations and probabilistic reasoning. I will sketch an outline of this option immediately below.

3.2 *Combining innateness with general learning*

Practitioners of TL often find themselves with two different hypotheses, call them F_1 and F_2, that seem equally capable of explaining the observed linguistic phenomena. F_1 and F_2 might come from entirely different theoretical frameworks, such as Combinatory Categorial Grammar and Minimalism for syntax or Optimality Theory and SPE for phonology, or they may constitute two refinements of the same broad framework. This has led to what Steedman and Baldridge (2011) have called a crisis in syntactic theory (though a similar problem arises in other subfields of TL, such as phonology and semantics): modern TL proposals are often meaningfully different in their essentials and yet comparably successful in accounting for the linguistic judgments of adult speakers. In order to choose between them, we need to look elsewhere.

One important source of evidence of this kind is the mapping from theories of competence to theories of processing, mediated by the competence hypothesis articulated by Miller and Chomsky (1963) and Chomsky (1965). This mapping has been used to argue for Lexical-Functional Grammar (over transformational grammars) by Bresnan and Kaplan (1982); for the flexible constituents endorsed by categorial grammars (over the rigid constituency of most other formalisms) by Steedman (1989); and for quantifier-raising (over *in situ* incorporation of quantifiers) by Hackl *et al.* (2012). I would like to suggest that combining CG with TL might provide another source of evidence of this kind, with a suitable mapping of UGs to CG learners (in Section 3.3 below I will argue that such a mapping is available by default through the principle of MDL). The shape of possible experiments to distinguish between F_1 and F_2 is as follows. Suppose one finds two properties, P_1 and P_2, that some languages have but some do not – so that learning will be involved – and that can co-exist in the same language. To take a phonological example, P_1 might be that a voiceless consonant like /*p*/ is aspirated in the beginning of a syllable while a voiced consonant like /*b*/ is not (as in English: *[p^h]at* vs. *[b]at*; note that this is a choice of English: Hindi can aspirate both /*p*/ and /*b*/, while French aspirates neither), and P_2 might be that vowels are lengthened before a voiced consonant but not before a voiceless consonant (again as in English: *t[a:]b* vs. *t[a]p*; again, this is a choice of English: French, for example, shows no such lengthening). In a syntactic example, P_1 might be that a subject can be dropped (as in Italian, but not in English) and P_2 might be that questions are marked by overt dislocation (again, as in Italian, but not in Japanese).

Given a CG mechanism M that seems cognitively plausible, we can now obtain two combinations, $M+F_1$ and $M+F_2$, and each combination can be run on a realistic corpus of child-directed speech. While F_1 and F_2 might both be capable of representing both P_1 and P_2, there might be a significant difference in how well the combinations $M+F_1$ and $M+F_2$ can learn the two and the order in which they do so. If this is the case, we now have a criterion for choosing between F_1 and F_2: whichever provides a better match with data from actual child language acquisition will receive support. Since M was proposed as a general-purpose learning mechanism and was not tailor made to handle either F_1 or F_2, such evidence can be taken seriously.

Experiments of this kind require researchers in each project to pay closer attention to work done in the other project than has usually been the case. Still, I think that they are a more productive – and, given current understanding, a more sensible – direction for future work on language and learning than further attempts to determine whether language is more innate than learned or vice versa.

3.3 *An argument for MDL*

I have tried to show why TL and CG can and should have a much closer relationship than they currently enjoy. In this section I will provide an argument that any explicit theory of UG already comes with the evaluation metric (or objective function) that forms the central component of a CG learner. Specifically, I will show how any explicit theory of UG translates into an MDL evaluation metric that allows the child to compare different possible hypotheses within the hypothesis space defined by UG. If correct, the discussion below points to bare MDL as our starting point in studying learning and as the linguist's M for comparing contenders for the correct theory of UG.

A theory of UG provides a set of possible grammars. Any of these can be the grammar of a competent speaker, who stores that grammar in memory and uses it to obtain an opinion about data. At the very least, then, assuming a theory of UG T with a set $\mathbb{G}$ of possible grammars commits us to the following assumptions:

1. A competent adult speaker has a grammar, $G \in \mathbb{G}$.
2. G is stored in memory.
3. G is used to parse inputs.

In order to make learning possible, we must allow a learner who currently represents G to also consider at least one other grammar G' and to switch from G to G' under certain conditions.[17] Of the very few properties that we can rely on to compare the two grammars in the general case, total storage space is a natural candidate, and one that accords well with the intuition behind MDL, which equates learning with compression. I therefore add the following two assumptions:

4. During language learning, a second grammar, $G' \in \mathbb{G}$ can be stored in memory and used to parse the input.
5. The memory size used to store G and its parse of the input can be compared to the memory size used to store G' and its parse of the input.

These assumptions amount to little more than saying that grammars can be used for parsing and that the overall description length of two grammars can be compared. My claim is that these assumptions already provide the language learner with an inherent learning mechanism: given an input D, the language learner searches through $\mathbb{G}$ for the grammar G for which the encoding of G (as defined by T) and of D (using G) is the shortest. By relying only on what the theory of UG under consideration is already committed to, this bare MDL learner offers a natural starting point for the study of learnability: alternatives in which the learner ignores the freely available MDL criterion and relies on some other mechanism instead should only be pursued

[17] Strictly speaking, maintaining more than one grammar is not always necessary. In particular, the learners proposed by Angluin (1982), Koshiba *et al.* (1997), Clark and Eyraud (2007), and Heinz (2010) all operate by considering just one grammar at a time and updating it as input comes along. All these learners, however, assume elaborate mechanisms for growing a grammar – usually tailor-made for the specific UGs they are designed to handle – that go well beyond the basic commitment to an explicit UG.

given evidence that the MDL null hypothesis is incorrect.[18, 19] The argument for bare MDL as the null hypothesis can be taken to support approaches in the literature that use MDL for learning, such as the works mentioned in Section 2.2.3, and in particular works such as de Marcken (1996) and Rasin and Katzir (2013) that use MDL not simply as a convenient heuristic but as the sole principle that maps an explicit UG to an evaluation metric.[20] Moreover, as mentioned in the introduction and discussed further in Section 3.2, the generality of

[18] To date, the literature has provided very little that bears directly on the empirical question of whether children use MDL as a criterion for comparing hypotheses during learning. On the other hand, several works have provided arguments – often in conflicting directions – regarding a possible role for description length more broadly in the learning process. In particular, Feldman (2000), extending the results of Shepard *et al.* (1961), provides evidence for the cognitive relevance of MDL by showing that description length is correlated with learning difficulty in concept learning (see also Feldman 2006 and Goodman *et al.* 2008). In the same vein, Moreton and Pater (2012a,b) review the literature on artificial grammar learning in phonology and conclude that description length is a central factor determining learning difficulty in this domain. On the other hand, Kurtz *et al.* (2013) point to a more nuanced pattern of difficulty in concept learning, and Moreton *et al.* (2014) provide evidence for correlating difficulty with factors other than description length, both in phonological learning and in concept learning. I will not attempt to relate such results about learning difficulty with the question of what evaluation criterion is used by the learner.

[19] Heinz and Idsardi (2013) note a lack of correlation between the complexity of finite-state machines for capturing certain patterns and potentially relevant language classes to which these patterns correspond. Based on this, Heinz and Idsardi suggest that MDL is not an appropriate learning criterion in phonology. Note, however, that the complexity of a grammar is only one part of the MDL criterion: the size of the description of the data given the grammar is just as important as the size of the grammar itself, and without taking it into account it is generally not possible to draw conclusions about the adequacy of the criterion. In addition, Heinz and Idsardi discuss the length of very specific representations – namely, the finite-state machines they use to describe the relevant patterns – and these representations do not correspond to any of the main grammatical formalisms for phonology. Given different representations, grammar size can change. Finally, it is hard to see how the possible correlation of language families with the description length for the best grammar (with or without taking the data into account) is a relevant consideration. The question is whether, given an appropriate representation scheme, the grammar that yields the shortest description in any particular situation is also the one that humans arrive at.

[20] For de Marcken (1996) MDL is a substitute for Structural Risk Minimiza-

the mapping from UGs to learners provides a framework in which theories of UG can be compared with respect to their predictions about learning.

4 A SIMPLE EXAMPLE

4.1 *Encoding*

To see how the mapping from theories of UG to bare MDL learners works, let us consider a naive theory of UG. This theory, call it T_1, allows any CFG to be represented by listing all the rules in some order, with a category #, which is not one of the terminals or nonterminals in the grammar, serving as a separator. Since T_1 only allows CFGs, it can list each rule unambiguously as the left-hand side followed by the list of the categories on the right-hand side.[21] T_1 marks the end of the grammar with an additional separator. For example, the grammar below will be listed as `ABA#ABC#A#BCD#...#EFG##`:

$$G := \begin{cases} A \rightarrow\ B\ A \\ A \rightarrow\ B\ C \\ A \rightarrow\ \epsilon \\ B \rightarrow\ C\ D \\ \vdots \\ E \rightarrow\ F\ G \end{cases}$$

We still need to specify how T_1 encodes the categories in the list. Sticking to simple-minded (and deliberately suboptimal) choices, we will use a fixed code-length scheme for the different categories, where each category will be encoded using $k = \lceil \lg(|Categories| + 1) \rceil$ bits:

tion, but it is still the sole contributor to the actual evaluation metric used by the learner. While de Marcken's focus is different from that of the present work – in particular, his emphasis on a specific representational framework that he develops can obscure the general applicability of MDL as an immediate CG learning criterion for any explicit UG – his work provides a particularly clear example of how pure MDL can fit in with a linguistically motivated UG.

[21] This particular choice of encoding individual rules would change in extensions of the learner beyond CFG, but the general point will not be affected. As long as the grammar can be stored and used for parsing, it can be encoded, and the encoding can be used in an MDL learner.

#	000
A	001
⋮	⋮
G	111

The number of bits per category, k, will have to be represented as well. We can do this by starting the code with a sequence of k 0's followed by a single 1, and by agreeing to treat $\underbrace{000}_{k}$ as #. Encoding the grammar above, then, will be $\underbrace{000}_{k}1\underbrace{001}_{k}\underbrace{010}_{k}\underbrace{001}_{k}\underbrace{000}_{k}\dots\underbrace{000}_{k}$, and the total length of encoding G will be $|G| \approx k \cdot [\sum_{r \in G} |r| + 1]$.

As for determining the encoding of the data, D, given G, T_1 first groups rules by their left-hand side, and then enumerates the expansions:

Rule	Code
$A \rightarrow BA$	00
$A \rightarrow BC$	01
$A \rightarrow \epsilon$	10
$B \rightarrow CD$	0
$B \rightarrow b$	1
$C \rightarrow c$	ϵ
⋮	⋮

Suppose now that G provides the following parse for D: $T = [_A[_B \dots]\ [_C \dots]]$. T_1 encodes this parse by traversing the tree in preorder, concatenating the code for each expansion choice given the left-hand side: $C(T) = C(A)C(A \rightarrow\ BC\ |A)C(\dots|B)\dots C(\dots|C)\dots$. In cases of ambiguity, T_1 takes the shortest encoding.

4.2 *Search*

Using the UG specified above as T_1, we can now take some input D and search for the grammar that minimizes the total description length of G and of the encoding of D given G. Any grammar G_0 that parses the input can serve as an initial hypothesis for the search. Moreover, G_0 provides a trivial upper bound on the size of the search, since the total description length provided by the target grammar is at most as large as that provided by G_0.

For T_1, there is a very simple grammar that is guaranteed to parse D and can serve as G_0. This grammar is what I will refer to as the *concatenation grammar for* Σ, where Σ is the alphabet in which D is written. If $\Sigma = \{\sigma_1, \ldots, \sigma_n\}$, the concatenation grammar for Σ is defined as follows:

$$G := \begin{cases} \gamma \rightarrow \ \sigma_1 \, \gamma \\ \vdots \\ \gamma \rightarrow \ \sigma_n \, \gamma \end{cases}$$

The concatenation grammar for Σ makes all texts of a certain length written in Σ equally easy to describe. It treats all symbols in all positions in D as equally good and therefore fails to capture any regularity other than the alphabet in which D is written. Consequently, it is only a good hypothesis for a random or near-random text. However, since it parses D it can serve as an initial hypothesis, and it provides an initial upper bound on the total description length using the target grammar.

Still, the bound provided by the concatenation grammar is huge, ruling out an exhaustive search. A greedy search is not likely to succeed, due to various local optima along the way. To address this problem, the search in the simulations below relies on Simulated Annealing (SA, Kirkpatrick *et al.* 1983), though I wish to emphasize that I am not trying to model the search procedure in humans, and my only claims concern the definition of the objective function, stated in terms of total description length. Indeed, it is quite possible that, even if they use the MDL criterion, humans will turn out to be incapable of exploring the search space effectively. If that is the case, the search component could make the learner – and with it the entire innate component – considerably more restrictive than suggested by the representational abilities of UG and by the MDL criterion.[22]

[22] The idea that a significant part of the restrictiveness of the innate component may be the result of constraints on learning has been pursued in the literature in various contexts. See Saffran (2003), Heinz (2007), and Heinz and Idsardi (2013), for example.

SA proceeds by comparing a current hypothesis to one of its neighbors, chosen at random, in terms of goodness, which in the present case is the total description length. That is, when a current hypothesis G is compared to one of its neighbors, G', $|G| + |D|G|$ is compared to $|G'| + |D|G'|$. If G' is better than G (that is, $|G'| + |D|G'| < |G| + |D|G|$), the search switches to G'. Otherwise, the choice of whether to switch to G' is made probabilistically and depends both on how much worse G' is and on a *temperature* parameter. The higher the temperature, the more likely the search is to switch from G to its bad neighbor G'. Similarly, the closer G and G' are in terms of overall description length, the more likely the search is to switch to G'. The temperature is initially set to a relatively high value, and it is gradually lowered as the search progresses, making the search increasingly greedy. The search ends when the temperature descends below a certain threshold.

For any grammar G, the neighbor grammar G' is generated as a variant of G in which one of the changes in the following list occurs:

1. An element, possibly a new nonterminal, is added to one of the rules.
2. An element is deleted from one of the rules.
3. A new rule of the form $X \rightarrow \epsilon$ is created for some category X.
4. A nonterminal in the right-hand side of a rule is replaced with its expansion according to some rule in the grammar.
5. A nonterminal X in the right-hand side of a rule is replaced with a new nonterminal Y, and a unit rule $Y \rightarrow X$ is added to the grammar.

The modification is chosen according to a uniform distribution over possible changes. All decisions in a given modification are made randomly as well (category for insertion, positions for insertion or deletion, etc.).

4.3 *Results*

In Section 4.1 above we saw the specification of T_1, a simple-minded CFG UG, and in Section 4.2 we saw the details of a search procedure that turns the MDL evaluation metric induced by T_1 into a learner. In this section we will see the results of running this learner on two extremely simple data sets: one that is the concatenation of words from an artificial lexicon and another that involves palindromes. Both tasks are loosely based on patterns that arise in natural language. The concatenation data set requires that the learner address the challenge of segmenting the input, a challenge solved by human learners, who are exposed to inputs that are for the most part unsegmented. The palindrome data set requires that the learner address the challenge of acquiring center embedding, a common pattern in natural languages. Despite this loose correspondence with natural language, the goal of the present section is not the realistic modeling of learning in humans – both T_1 and the data sets are far too simplistic to be informative in this respect – but rather to show how a bare MDL learner induced by an explicit UG operates, and how the representational abilities of the UG in question guide the search for the best hypothesis given the data.

4.3.1 Segmentation

The first data set is based on the one described by Saffran *et al.* (1996). In Saffran *et al.*'s experiment, in which a text was generated by the random concatenation of elements from an artificial vocabulary consisting of the items `pabiku, golatu, daropi, tibudo`. This text was turned into speech using a synthesizer that produced a stream of speech with flat intonation and no word breaks. Eight-month old infants were exposed to this stream, and after two minutes (= 180 words = 1080 segments) they were able to distinguish between words (e.g. `pabiku`) and non-word sequences that appear in the text (e.g. `bikuda`).[23] Here are sample snapshots from the learning process using an input that is only 300 segments long (compared to 1080 in the original experiment), using an initial temperature of 15 and a maximum grammar-length of 200 bits. The first step, as explained above, is a concatenation grammar, which captures no regularities:[24]

[23] The text used by Saffran *et al.* (1996) was subject to the additional requirement that no word can repeat itself. In the text that I used, repetitions are not prohibited. As far as I can tell, this does not affect the point made here.

[24] In the results reported here, the step in the search appears as the subscript of G; γ is the seed category; and numbered categories are non-terminal categories that are hypothesized by the learner during the search.

$$\begin{array}{lll} G_0: & \gamma \rightarrow k\ \gamma & \gamma \rightarrow i\ \gamma \\ & \gamma \rightarrow o\ \gamma & \gamma \rightarrow u\ \gamma \\ & \gamma \rightarrow d\ \gamma & \gamma \rightarrow p\ \gamma \\ & \gamma \rightarrow a\ \gamma & \gamma \rightarrow g\ \gamma \\ & \gamma \rightarrow r\ \gamma & \gamma \rightarrow b\ \gamma \\ & \gamma \rightarrow l\ \gamma & \gamma \rightarrow t\ \gamma \end{array}$$

Grammar length: 126, Encoding length: 1200, Energy: 1326.0

After a thousand steps, we already have `ro` from `daropi`, `la` and `go` from `golatu`, and `ku` from `pabiku`:

$$\begin{array}{lll} G_{1000}: & d \rightarrow o & \gamma \rightarrow d\ \gamma \\ & \gamma \rightarrow \gamma & \gamma \rightarrow u\ \gamma\ d \\ & a \rightarrow & \gamma \rightarrow o\ \gamma\ g \\ & \gamma \rightarrow t\ \gamma & \gamma \rightarrow l\ a\ \gamma\ i \\ & \gamma \rightarrow r\ o\ \gamma & \gamma \rightarrow g\ o\ \gamma\ p \\ & \gamma \rightarrow i\ \gamma\ t & \gamma \rightarrow p\ \gamma\ d \\ & l \rightarrow u\ i & \gamma \rightarrow k\ u\ \gamma\ b \\ & \gamma \rightarrow a\ \gamma & r \rightarrow \\ & \gamma \rightarrow b\ \gamma & \end{array}$$

Grammar length: 192, Encoding length: 1023, Energy: 1215.0

As we proceed, more and more parts of the underlying vocabulary are discovered. Here, at the final step, we have all the words:

$$\begin{array}{lll} G_{100000}: & 5144 \rightarrow t\ i\ b\ u\ d\ o\ 5144 & 5144 \rightarrow p\ a\ b\ i\ k\ u\ 5144 \\ & 5144 \rightarrow g\ o\ l\ a\ t\ u\ 5144\ r & 5144 \rightarrow d\ a\ r\ o\ p\ i\ 5144 \end{array}$$

Grammar length: 97, Encoding length: 100, Energy: 197.0

The results presented above show rules that correspond straightforwardly to the lexicon that was used to generate the input and thus reflect the correct segmentation of the input, based on its statistical regularities. Crucially, though, the theory of UG presented as T_1 in Section 4.1 is not aware of the tasks of segmentation and lexicon induction, and it does not represent probabilities in its rules. Consequently, the bare MDL learner for T_1 is not aware of these notions either. It arrives at the correct segmentation as a by-product of its general search for the best grammar given the input.

4.3.2 Palindromes

For our second simulation, along the lines of Horning's paradigm, we will use an input that exhibits nested dependencies. Such dependencies are common in natural language: they are present in the nesting of object-extracted relative clauses in English, for example, as well as in the basic structure of verb-argument dependencies in German clauses. It has been suggested by Fitch and Hauser (2004) that humans acquire such patterns in experiments of artificial-language learning, though the experiment and the claim remain controversial (see Perruchet and Rey 2005, among others).[25]

In the nesting data set I will use a segmented input. We can specify the learner's goal when presented with a segmented input sequence to be the minimization of the sum of the grammar length and the sum of the encoding lengths for each element in the sequence.[26] At least in simple cases, the learner successfully identifies the generating grammar from an input presented in this way. Following are several snapshots from a run on an input that consists of 200 even-lengthed palindromes over the alphabet $\Sigma = \{a, b, c\}$ (the sequence reported here starts as $cccabaccabaccc, cbbc, bcccccb, aa, aabbaa, \ldots$; for performance purposes, the learner cannot see past the first 25 characters of each element in the sequence):

$$G_0: \quad \gamma \rightarrow a\ \gamma \qquad \gamma \rightarrow c\ \gamma$$
$$\gamma \rightarrow b\ \gamma$$

Grammar length: 19, Encoding length: 2314, Energy: 2333.0

[25] The palindrome language is a member of certain interesting infinite classes that can also be learned under the demanding criterion of *iitl*, as shown by Koshiba *et al.* (1997).

[26] Note, however, that the learner treats its input as the prefix of a possibly infinite text rather than a complete element in the language. I will not discuss this issue.

$$\begin{array}{lll} G_{1400}: & \gamma \rightarrow c\ \gamma & \gamma \rightarrow a\ \gamma\ b\ \gamma \\ & \gamma \rightarrow c & \gamma \rightarrow b\ \gamma\ c\ b\ \gamma \end{array}$$

Grammar length: 32, Encoding length: 2122, Energy: 2154.0

$$\begin{array}{lll} G_{2800}: & 209 \rightarrow c\ 209 & 209 \rightarrow a\ 209 \\ & 209 \rightarrow b\ 209\ b\ c\ c\ 209\ c\ b\ 209\ a & 209 \rightarrow \end{array}$$

Grammar length: 35, Encoding length: 2154, Energy: 2189.0

$$\begin{array}{lll} G_{4200}: & 371 \rightarrow a\ 371\ a & 371 \rightarrow \\ & 371 \rightarrow b\ 371\ b & 371 \rightarrow c\ 371\ c \end{array}$$

Grammar length: 27, Encoding length: 1480, Energy: 1507.0

G_{4200} is already the correct grammar (371 is the arbitrary category label of what would usually be written as S). Similar results were obtained with other simple CFGs, such as $a^n b^n$.

5 DISCUSSION

I set out to bring TL theories of UG and CG theories of learning into closer contact. I reviewed some of the central arguments within each discipline for and against rich UGs and for and against learning, concluding that linguists' notions of rich UGs are well-founded, but that cognition-general learning approaches are viable as well. Differently from what is often suggested in the literature, I argued that the two can and should co-exist and support each other. Specifically, I used the observation that any theory of UG provides a learning criterion – the total memory space used to store a grammar and its encoding of the input – that supports an MDL evaluation metric that can serve as the central component of a CG learner. This mapping from theories of UG to learners maintains a minimal ontological commitment: the learner for a particular theory of UG uses only what that theory already requires to account for linguistic competence in adults. I suggested that such learners should be our null hypothesis regarding the child's learning mechanism, and that furthermore, the mapping from theories of UG to learners provides a framework for comparing theories of UG.

ACKNOWLEDGEMENTS

I thank Adam Albright, Asaf Bachrach, Bob Berwick, Michael Collins, Danny Fox, Tova Friedman, Arnon Lotem, Alec Marantz, Andrew Nevins, Tim O'Donnell, Amy Perfors, Ezer Rasin, Raj Singh, Donca Steriade, Josh Tenenbaum, and the audiences at MIT, Tel Aviv University, Ben Gurion University, Hebrew University, and the Cornell Workshop on Grammar Induction, as well as the reviewers and editors of JLM. This work has been supported by ISF grant 187/11.

REFERENCES

Klaus ABELS and Ad NEELEMAN (2010), Nihilism masquerading as progress, *Lingua*, 120(12):2657–2660, ISSN 0024-3841.

Dana ANGLUIN (1980), Inductive inference of formal languages from positive data, *Information and Control*, 45:117–135.

Dana ANGLUIN (1982), Inference of Reversible Languages, *Journal of the Association for Computing Machinery*, 29(3):741–765.

Dana ANGLUIN (1988), Identifying Languages from Stochastic Examples, Technical Report 614, Yale University.

Richard N. ASLIN, Jenny R. SAFFRAN, and Elissa L. NEWPORT (1998), Computation of conditional probability statistics by 8-month old infants, *Psychological Science*, 9:321–324.

Carl L. BAKER (1979), Syntactic theory and the projection problem, *Linguistic Inquiry*, 10(4):533–581.

Eleanor BATCHELDER (2002), Bootstrapping the Lexicon: A Computational Model of Infant Speech Segmentation, *Cognition*, 83:167–206.

Michael BECKER, Nihan KETREZ, and Andrew NEVINS (2011), The Surfeit of the Stimulus: Analytic Biases Filter Lexical Statistics in Turkish Laryngeal Alternations, *Language*, 87(1):84–125.

Robert C. BERWICK (1982), *Locality Principles and the Acquisition of Syntactic Knowledge*, Ph.D. thesis, MIT, Cambridge, MA.

Robert C. BERWICK, Paul PIETROSKI, Beracah YANKAMA, and Noam CHOMSKY (2011), Poverty of the Stimulus Revisited, *Cognitive Science*, 35(7):1207–1242, ISSN 1551-6709.

Paul BOERSMA and Bruce HAYES (2001), Empirical Tests of the Gradual Learning Algorithm, *Linguistic Inquiry*, 32:45–86.

Luca BONATTI, Marcela PEÑA, Marina NESPOR, and Jacques MEHLER (2005), Linguistic Constraints on Statistical Computations, *Psychological Science*, 16(6):451–459.

Martin D. S. BRAINE (1971), On Two Types of Models of the Internalization of Grammars, in D. J. SLOBIN, editor, *The Ontogenesis of Grammar*, pp. 153–186, Academic Press.

Michael BRENT (1999), An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery, *Computational Linguistics*, 34(1–3):71–105.

Michael BRENT and T. CARTWRIGHT (1996), Distributional Regularity and Phonotactic Constraints are useful for Segmentation, *Cognition*, 61:93–125.

Joan BRESNAN and Ronald M. KAPLAN (1982), Grammars as Mental Representations of Language, in *The Mental Representation of Grammatical Relations*, MIT Press.

Roger BROWN and Camille HANLON (1970), Derivational Complexity and the Order of Acquisition of Child Speech, in J. R. HAYES, editor, *Cognition and the Development of Language*, pp. 11–53, Wiley, New York.

Gregory J. CHAITIN (1966), On the Length of Programs for Computing Finite Binary Sequences, *Journal of the ACM*, 13:547–569.

Nancy Chih-Lin CHANG (2008), *Constructing grammar: A computational model of the emergence of early constructions*, Ph.D. thesis, University of California, Berkeley, CA.

Moses CHARIKAR, Eric LEHMAN, Ding LIU, Rina PANIGRAHY, Manoj PRABHAKARAN, Amit SAHAI, and Abhi SHELAT (2005), The smallest grammar problem, *Information Theory, IEEE Transactions on*, 51(7):2554–2576.

Nick CHATER and Paul VITÁNYI (2007), 'Ideal Learning' of Natural Language: Positive Results about Learning from Positive Evidence, *Journal of Mathematical Psychology*, 51:135–163.

Stanley CHEN (1996), *Building Probabilistic Models for Natural Language*, Ph.D. thesis, Harvard University, Cambridge, MA.

Noam CHOMSKY (1965), *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

Noam CHOMSKY (1981), *Lectures on Government and Binding*, Foris, Dordrecht.

Noam CHOMSKY and Morris HALLE (1968), *The Sound Pattern of English*, Harper and Row Publishers, New York.

Morten CHRISTIANSEN, Joseph ALLEN, and Mark SEIDENBERG (1998), Learning to Segment Speech using Multiple Cues: A Connectionist Model, *Language and Cognitive Processes*, 13(2/3):221–268.

Alexander CLARK (2001), *Unsupervised Language Acquisition: Theory and Practice*, Ph.D. thesis, University of Sussex.

Alexander CLARK and Rémi EYRAUD (2007), Polynomial identification in the limit of context-free substitutable languages, *Journal of Machine Learning Research*, 8:1725–1745.

Alexander CLARK and Shalom LAPPIN (2011), *Linguistic Nativism and the Poverty of the Stimulus*, Wiley-Blackwell.

Robin CLARK and Ian ROBERTS (1993), A computational model of language learnability and language change, *Linguistic Inquiry*, 24(2):299–346.

Stephen CRAIN, Drew KHLENTZOS, and Rosalind THORNTON (2010), Universal Grammar versus language diversity, *Lingua*, 120(12):2668–2672, ISSN 0024-3841.

Stephen CRAIN and Paul PIETROSKI (2002), Why Language Acquisition is a Snap, *The Linguistic Review*, 19:163–183.

Carl DE MARCKEN (1996), *Unsupervised Language Acquisition*, Ph.D. thesis, MIT, Cambridge, MA.

François DELL (1981), On the learnability of optional phonological rules, *Linguistic Inquiry*, 12(1):31–37.

Łukasz DĘBOWSKI (2011), On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts, *Information Theory, IEEE Transactions on*, 57(7):4589–4599, ISSN 0018-9448, doi:10.1109/TIT.2011.2145170.

Mike DOWMAN (2007), Minimum Description Length as a Solution to the Problem of Generalization in Syntactic Theory, ms., University of Tokyo, Under review.

Fred C. DYER and Jeffrey A. DICKINSON (1994), Development of sun compensation by honeybees: How partially experienced bees estimate the sun's course, *Proceedings of the National Academy of Sciences*, 91(10):4471–4474.

Ansgar ENDRESS, Ghislaine DEHAENE-LAMBERTZ, and Jacques MEHLER (2007), Perceptual Constraints and the Learnability of Simple Grammars, *Cognition*, 105(3):577–614.

Ansgar ENDRESS, Marina NESPOR, and Jacques MEHLER (2009), Perceptual and Memory Constraints on Language Acquisition, *Trends in Cognitive Sciences*, 13(8):348–353.

Ansgar D. ENDRESS and Luca L. BONATTI (2013), Single vs. multiple mechanism models of artificial grammar learning, under review.

Ansgar D. ENDRESS and Jacques MEHLER (2009), Primitive computations in speech processing, *The Quarterly Journal of Experimental Psychology*, 62(11):2187–2209.

Ansgar D ENDRESS and Jacques MEHLER (2010), Perceptual constraints in phonotactic learning, *Journal of experimental psychology. Human perception and performance*, 36(1):235–250.

Nicholas EVANS and Stephen LEVINSON (2009), The Myth of Language Universals: Language Diversity and its Importance for Cognitive Science, *Behavioral and Brain Sciences*, 32:429–492.

Olga FEHER, Haibin WANG, Sigal SAAR, Partha P. MITRA, and Ofer TCHERNICHOVSKI (2009), De novo establishment of wild-type song culture in the zebra finch, *Nature*, 459(7246):564–568.

Jacob FELDMAN (2000), Minimization of Boolean complexity in human concept learning, *Nature*, 407(6804):630–633.

Jacob FELDMAN (2006), An algebra of human concept learning, *Journal of Mathematical Psychology*, 50(4):339–368, ISSN 0022-2496.

W. Tecumseh FITCH and Marc D. HAUSER (2004), Computational constraints on syntactic processing in a nonhuman primate, *Science*, 303(5656):377–380.

Stephani FORAKER, Terry REGIER, Naveen KHETARPAL, Amy PERFORS, and Joshua TENENBAUM (2009), Indirect Evidence and the Poverty of the Stimulus: The Case of Anaphoric One, *Cognitive Science*, 33(2):287–300, ISSN 1551-6709.

John GARCIA, Walter HANKINS, and Kenneth RUSINIAK (1974), Behavioral Regulation of the Milieu Interne in Man and Rat, *Science*, 185(4154):824–831.

Edward GIBSON and Kenneth WEXLER (1994), Triggers, *Linguistic Inquiry*, 25(3):407–454.

E. Mark GOLD (1967), Language Identification in the Limit, *Information and Control*, 10:447–474.

John GOLDSMITH (2001), Unsupervised Learning of the Morphology of a Natural Language, *Computational Linguistics*, 27(2):153–198.

Noah D. GOODMAN, Joshua B. TENENBAUM, Jacob FELDMAN, and Thomas L. GRIFFITHS (2008), A Rational Analysis of Rule-Based Concept Learning, *Cognitive Science*, 32(1):108–154.

Thomas GRIFFITHS and Joshua TENENBAUM (2006), Optimal Predictions in Everyday Cognition, *Psychological Science*, 17(9):767–773.

Peter GRÜNWALD (1996), A Minimum Description Length Approach to Grammar Inference, in G. S. S. WERMTER and E. RILOFF, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Springer Lecture Notes in Artificial Intelligence, pp. 203–216, Springer.

Martin HACKL, Jorie KOSTER-HALE, and Jason VARVOUTIS (2012), Quantification and ACD: Evidence from Real-Time Sentence Processing, *Journal of Semantics*, 29(2):145–206.

Daniel HARBOUR (2011), Mythomania? Methods and morals from 'The Myth of Language Universals', *Lingua*, 121(12):1820–1830, ISSN 0024-3841.

Zellig S. HARRIS (1955), From Phoneme to Morpheme, *Language*, 31(2):190–222.

Jeffrey HEINZ (2007), *The Inductive Learning of Phonotactic Patterns*, Ph.D. thesis, University of California, Los Angeles.

Jeffrey HEINZ (2010), String Extension Learning, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 897–906.

Jeffrey HEINZ and William IDSARDI (2013), What Complexity Differences Reveal About Domains in Language, *Topics in cognitive science*, 5(1):111–131.

Laurence HORN (1972), *On the Semantic Properties of the Logical Operators in English*, Ph.D. thesis, UCLA.

Laurence HORN (2011), Histoire d'*O: Lexical Pragmatics and the Geometry of Opposition, in Jean-Yves BÉZIAU and Gilbert PAYETTE, editors, *The Square of Opposition: A General Framework for Cognition*, pp. 383–416, Peter Lang.

James HORNING (1969), *A Study of Grammatical Inference*, Ph.D. thesis, Stanford.

Anne S. HSU and Nick CHATER (2010), The Logical Problem of Language Acquisition: A Probabilistic Perspective, *Cognitive Science*, 34(6):972–1016, ISSN 1551-6709.

Anne S. HSU, Nick CHATER, and Paul M.B. VITÁNYI (2011), The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis, *Cognition*, 120(3):380 – 390, ISSN 0010-0277.

Tim HUNTER and Jeffrey LIDZ (2013), Conservativity and Learnability of Determiners, *Journal of Semantics*, 30(3):315–334.

Elizabeth K. JOHNSON and Peter W. JUSCZYK (2001), Word Segmentation by 8-Month Olds: When Speech Cues count more than Statistics, *Journal of Memory and Language*, 44:548–567.

Shyam KAPUR (1991), *Computational learning of languages*, Ph.D. thesis, Cornell University, Ithaca, NY.

Roni KATZIR and Raj SINGH (2013), Constraints on the Lexicalization of Logical Operators, *Linguistics and Philosophy*, 36(1):1–29.

John C. KIEFFER and En-hui YANG (2000), Grammar-based codes: a new class of universal lossless source codes, *Information Theory, IEEE Transactions on*, 46(3):737–754, ISSN 0018-9448, doi:10.1109/18.841160.

Simon KIRBY (2000), Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners, *The evolutionary emergence of language: Social function and the origins of linguistic form*, pp. 303–323.

Simon KIRBY (2002), Learning, bottlenecks and the evolution of recursive syntax, *Linguistic evolution through language acquisition: Formal and computational models*, pp. 173–203.

Simon KIRBY, Kenny SMITH, and Henry BRIGHTON (2004), From UG to Universals., *Studies in Language*, 28(3):587–607.

Scott KIRKPATRICK, C. Daniel GELATT, and Mario P. VECCHI (1983), Optimization by Simulated Annealing, *Science*, 220(4598):671–680.

Andrei Nikolaevic KOLMOGOROV (1965), Three Approaches to the Quantitative Definition of Information, *Problems of Information Transmission (Problemy Peredachi Informatsii)*, 1:1–7, republished as Kolmogorov (1968).

Andrei Nikolaevic KOLMOGOROV (1968), Three Approaches to the Quantitative Definition of Information, *International Journal of Computer Mathematics*, 2:157–168.

Takeshi KOSHIBA, Erkki MÄKINEN, and Yuji TAKADA (1997), Learning deterministic even linear languages from positive examples, *Theoretical Computer Science*, 185(1):63 – 79, ISSN 0304-3975.

Kenneth J KURTZ, Kimery R LEVERING, Roger D STANTON, Joshua ROMERO, and Steven N MORRIS (2013), Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961), *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2):552–572.

Julie Anne LEGATE and Charles YANG (2002), Empirical Re-assessment of Stimulus Poverty Arguments, *The Linguistic Review*, 19(151–162).

Abraham LEMPEL and Jacob ZIV (1976), On the Complexity of Finite Sequences, *IEEE Transactions on Information Theory*, 22(1):75–81, ISSN 0018-9448.

Stephen C. LEVINSON and Nicholas EVANS (2010), Time for a sea-change in linguistics: Response to comments on 'The Myth of Language Universals', *Lingua*, 120(12):2733–2758, ISSN 0024-3841.

Ming LI and Paul VITÁNYI (1997), *An Introduction to Kolmogorov Complexity and its Applications*, Springer Verlag, Berlin, 2nd edition.

Jeffrey LIDZ, Sandra WAXMAN, and Jennifer FREEDMAN (2003), What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months, *Cognition*, 89:B65–B73.

Giorgio MAGRI (2013), The Complexity of Learning in Optimality Theory and Its Implications for the Acquisition of Phonotactics, *Linguistic Inquiry*, 44(3):433–468.

Gary F. MARCUS (1993), Negative Evidence in Language Acquisition, *Cognition*, 46:53–85.

Gary F. MARCUS (2000), Pabiku and Ga Ti Ga: Two Mechanisms Infants Use to Learn about the World, *Current Directions in Psychological Science*, 9:145–147.

Lisa MATTHEWSON (2012), On How (Not) to Uncover Cross-Linguistic Variation, in *Proceedings of NELS 42*.

Sven MATTYS, Peter W. JUSCZYK, Paul LUCE, and James L. MORGAN (1999), Phonotactic and Prosodic Effects on Word Segmentation in Infants, *Cognitive Psychology*, 38:465–494.

George MILLER and Noam CHOMSKY (1963), Finitary Models of Language Users, in R. Duncan LUCE, Robert R. BUSH, and Eugene GALANTER, editors, *Handbook of Mathematical Psychology*, volume 2, pp. 419–491, Wiley, New York, NY.

Elliott MORETON (2008), Analytic Bias and Phonological Typology, *Phonology*, 25:83–127.

Elliott MORETON and Joe PATER (2012a), Structure and Substance in Artificial-phonology Learning, Part I: Structure, *Language and Linguistics Compass*, 6(11):686–701, ISSN 1749-818X.

Elliott MORETON and Joe PATER (2012b), Structure and Substance in Artificial-Phonology Learning, Part II: Substance, *Language and Linguistics Compass*, 6(11):702–718, ISSN 1749-818X.

Elliott MORETON, Joe PATER, and Katya PERTSOVA (2014), Phonological concept learning, ms., Under review.

Craig NEVILL-MANNING and Ian WITTEN (1997), Compression and Explanation using Hierarchical Grammars, *The Computer Journal*, 40(2 and 3):103–116.

Partha NIYOGI (2006), *The Computational Nature of Language and Learning*, MIT Press.

Partha NIYOGI and Robert C. BERWICK (1996), A Language Learning Model for Finite Parameter Spaces, *Cognition*, 61:161–193.

Partha NIYOGI and Robert C. BERWICK (1997), Evolutionary consequences of language learning, *Linguistics and Philosophy*, 20(6):697–719.

Partha NIYOGI and Robert C. BERWICK (2009), The proper treatment of language acquisition and change in a population setting, *Proceedings of the National Academy of Sciences*, 106:10124–10129.

Luca ONNIS, Matthew ROBERTS, and Nick CHATER (2002), Simplicity: A Cure for Overgeneralization in Language Acquisition?, in W. D. GRAY and C. D. SHUNN, editors, *Proceedings of the 24th Annual Conference of the Cognitive Society*, London.

Miles OSBORNE and Ted BRISCOE (1997), Learning Stochastic Categorial Grammars, in *Proceedings of CoNLL*, pp. 80–87.

Daniel N. OSHERSON, Michael STOB, and Scott WEINSTEIN (1984), Learning Theory and Natural Language, *Cognition*, 17:1–28.

Daniel N. OSHERSON, Michael STOB, and Scott WEINSTEIN (1986), *Systems that learn*, MIT Press, Cambridge, Massachusetts.

Marcela PEÑA, Luca BONATTI, Marina NESPOR, and Jacques MEHLER (2002), Signal-Driven Computations in Speech Processing, *Science*, 298:604–607.

Amy PERFORS, Joshua TENENBAUM, and Terry REGIER (2011), The Learnability of Abstract Syntactic Principles, *Cognition*, 118:306–338.

Pierre PERRUCHET and Arnaud REY (2005), Does the Mastery of Center-Embedded Linguistic Structures Distinguish Humans from Nonhuman Primates?, *Psychonomic Bulletin and Review*, 12(2):307–313.

Steven T. PIANTADOSI and Edward GIBSON (2013), Quantitative Standards for Absolute Linguistic Universals, *Cognitive Science*, pp. n/a–n/a, ISSN 1551-6709, doi:10.1111/cogs.12088.

Leonard PITT (1989), Probabilistic Inductive Inference, *Journal of the ACM*, 36(2):383–433.

Alan PRINCE and Paul SMOLENSKY (1993), Optimality Theory: Constraint Interaction in Generative Grammar, Technical report, Rutgers University, Center for Cognitive Science.

Ezer RASIN and Roni KATZIR (2013), On evaluation metrics in Optimality Theory, ms., MIT and TAU (submitted).

Eric REULAND and Martin EVERAERT (2010), Reaction to: The Myth of Language Universals and cognitive science"—Evans and Levinson's cabinet of curiosities: Should we pay the fee?, *Lingua*, 120(12):2713–2716, ISSN 0024-3841.

Jorma RISSANEN (1978), Modeling by Shortest Data Description, *Automatica*, 14:465–471.

Jorma RISSANEN and Eric Sven RISTAD (1994), Language Acquisition in the MDL Framework, in *Language computations: DIMACS Workshop on Human Language, March 20-22, 1992*, p. 149, Amer Mathematical Society.

John R. ROSS (1967), *Constraints on Variables in Syntax*, Ph.D. thesis, MIT, Cambridge, MA.

Jenny R. SAFFRAN (2003), Statistical Language Learning: Mechanisms and Constraints, *Current Directions in Psychological Science*, 12(4):110–114.

Jenny R. SAFFRAN, Elissa L. NEWPORT, and Richard N. ASLIN (1996), Statistical learning by 8-month old infants, *Science*, 274:1926–1928.

Wendy SANDLER, Irit MEIR, Carol PADDEN, and Mark ARONOFF (2005), The emergence of grammar: Systematic structure in a new language, *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2661–2665.

Uli SAUERLAND and Jonathan BOBALJIK (2013), Syncretism Distribution Modeling: Accidental Homophony as a Random Event, in *Proceedings of GLOW in Asia IX 2012*.

Ann SENGHAS, Sotaro KITA, and Asli ÖZYÜREK (2004), Children Creating Core Properties of Language: Evidence from an Emerging Sign Language in Nicaragua, *Science*, 305(5691):1779–1782.

Roger N SHEPARD, Carl I HOVLAND, and Herbert M JENKINS (1961), Learning and memorization of classifications, *Psychological Monographs: General and Applied*, 75(13):1–42.

Kenny SMITH, Simon KIRBY, and Henry BRIGHTON (2003), Iterated learning: A framework for the emergence of language, *Artificial Life*, 9(4):371–386.

Kirk H. SMITH (1966), Grammatical Intrusions in the Recall of Structured Letter Pairs: Mediated Transfer or Position Learning?, *Journal of Experimental Psychology*, 72(4):580–588.

David M. SOBEL, Joshua B. TENENBAUM, and Alison GOPNIK (2004), Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers, *Cognitive science*, 28(3):303–333.

Ray J. SOLOMONOFF (1964), A formal theory of inductive inference, parts I and II, *Information and Control*, 7(1 & 2):1–22, 224–254.

Ray J. SOLOMONOFF (1978), Complexity-Based Induction Systems: Comparisons and Convergence Theorems, *IEEE Transactions on Information Theory*, 24(4):422–432.

Ray J. SOLOMONOFF (2008), Algorithmic Probability: Theory and Applications, in Frank EMMERT-STREIB and Matthias DEHMER, editors, *Information Theory and Statistical Learning*, pp. 1–23, Springer.

Mark STEEDMAN (1989), Grammar, Interpretation, and Processing from the Lexicon, in William MARSLEN-WILSON, editor, *Lexical Representation and Process*, pp. 463–504, MIT Press.

Mark STEEDMAN and Jason BALDRIDGE (2011), Combinatory Categorial Grammar, in Robert BORSLEY and Kersti BÖRJARS, editors, *Non-Transformational Syntax*, chapter 5, pp. 181–224, Blackwell.

Andreas STOLCKE (1994), *Bayesian Learning of Probabilistic Language Models*, Ph.D. thesis, University of California at Berkeley, Berkeley, California.

Bruce TESAR and Paul SMOLENSKY (1998), Learnability in Optimality Theory, *Linguistic Inquiry*, 29(2):229–268.

Harry TILY and T. Florian JAEGER (2011), Complementing quantitative typology with behavioral approaches: Evidence for typological universals, *Linguistic Typology*, 15(2):497–508.

Anand VENKATARAMAN (2001), A Statistical Model for Word Discovery in Transcribed Speech, *Computational Linguistics*, 27(3):351–372.

Christopher S. WALLACE and David M. BOULTON (1968), An Information Measure for Classification, *Computer Journal*, 11(2):185–194.

Kenneth WEXLER and Peter W. CULICOVER (1980), *Formal Principles of Language Acquisition*, MIT Press, Cambridge, MA.

Colin WILSON (2006), Learning Phonology with Substantive Bias: An Experimental and Computational Study of Velar Palatalization, *Cognitive Science*, 30(5):945–982.

Charles D. YANG (2002), *Knowledge and learning in natural language*, Oxford University Press.

Charles D. YANG (2004), Universal Grammar, statistics or both?, *Trends in Cognitive Sciences*, 8(10):451–456.

Charles D. YANG (2010), Three Factors in Language Variation, *Lingua*, 120:1160–1177.

Ryo YOSHINAKA (2011), Efficient learning of multiple context-free languages with multidimensional substitutability from positive data, *Theoretical Computer Science*, 412(19):1821–1831, ISSN 0304-3975, doi:http://dx.doi.org/10.1016/j.tcs.2010.12.058.

Willem ZUIDEMA (2003), How the Poverty of the Stimulus Solves the Poverty of the Stimulus, in Suzanna BECKER, Sebastian THRUN, and Klaus OBERMAYER, editors, *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02)*, pp. 51–58.

A proof-theoretic semantics for contextual domain restriction

Nissim Francez
Computer Science Department,
The Technion-IIT, Haifa, Israel

ABSTRACT

The paper presents a proof-theoretic semantics account of contextual domain restriction for quantified sentences in a fragment of English. First, the technique is exemplified in the more familiar first-order logic, and in its restricted quantification variant. Then, a proof-theoretic semantics for the NL fragment is reviewed, and extended to handling contextual domain restriction. The paper addresses both the descriptive facet of the problem, deriving meaning relative to a context, as well as the fundamental aspect, defining explicitly a context (suitable for quantifier domain restriction), and specifying what it is about such a context that brings about the variation of meaning due to it.

The paper argues for the following principle (*the context incorporation principle, CIP*): for every quantified sentence S depending on a context c, there exists a sentence S', the meaning of which is independent of c, s.t. the contextually restricted meaning of S is equal to the meaning of S'. Thus, the effect of a context can always be internalized. The current model-theoretic accounts of contextual domain restriction do not satisfy CIP, in that they imply intersection of some extension with an arbitrary subset of the domain, that need not be the denotation of any NL-expression.

Keywords:
proof-theoretic semantics,
contextual domain restriction

Journal of Language Modelling Vol 2, No 2 (2014), pp. 249–283

1 INTRODUCTION

> The problem of context dependence is the problem of explaining how context contributes to interpretation …

write Stanley and Szabó (2000), who discuss a variety of special cases of the general problem of meaning variation with context. The purpose of this paper is to provide a *proof-theoretic semantics (PTS)*[1] (see below) for a special case of the general context dependence problem, namely *quantifier domain restriction (QDR)*. It may well be the case that the proof-theoretic interpretation of other kinds of expressions with contextually varying meanings will require different proof-theoretic techniques than the one used here. I focus on the QDR-problem as it fits naturally into the fragment of natural language (NL) for which a PTS has been proposed before (Francez and Dyckhoff 2010; Francez *et al.* 2010; Francez and Ben-Avi 2014). The QDR-problem has a rich history (see Stanley and Szabó 2000, for references to earlier work), all carried out under the *model-theoretic semantics (MTS)* theory of meaning.

Before turning to the main semantic issue itself, I would like to recapitulate the highlights of the PTS and MTS approaches as theories of meaning. Proof-theoretic semantics is a challenging way for defining meaning, an alternative to the prevailing model-theoretic semantics, the latter equating meaning with providing *truth conditions* (in arbitrary models).[2] The MTS approach has been criticized by several philosophers of logic and language (most notably, Dummett 1993, Prawitz 2006, Brandom 2000, Tennant 1997, and many more) as an inappropriate theory of meaning. I omit here a more detailed discussion of this criticism, often occupying full books, as justifying the approach is not the topic here. A more condensed presentation of this criticism and motivating advantages of PTS can be found in the introduction sections of (Francez and Dyckhoff 2010) and (Francez and

[1] A general introductory overview of PTS can be found in an entry of *The Stanford Encyclopaedia of Philosophy*, `http://plato.stanford.edu/entries/proof-theoretic-semantics/`. Concrete references are given in the paper where appropriate.

[2] There is also a variant of MTS called *Dynamic Semantics*, which view meaning as *updates* of assignments. It also depends on models, entities, reference, extension, etc.

Ben-Avi 2014). Initially, since the work of Gentzen (1969), PTS was conceived as a meaning-theory for logic. Recently, however, PTS has been advocated also for providing the semantics of (an extensional fragment of) NL in (Francez and Dyckhoff 2010), (Francez *et al.* 2010) and (Francez and Ben-Avi 2014), in contrast to the MTS approach dominant in NL formal semantics ever since Montague's seminal work.

I recapitulate the essence of the PTS proposal:

- *For (affirmative) sentences, replace the received approach of taking their meanings as* **truth conditions** *(in arbitrary models) by an approach taking meanings to consist of* **canonical derivability conditions** *(from suitable assumptions)*. This involves a dedicated proof system in natural deduction (ND) form, on which the derivability conditions are based (canonicity is explained below). In a sense, the proof system should reflect the use of the sentences in the fragment, and should allow recovering pre-theoretic properties of the meanings of these sentences such as entailment and assertability conditions. The essentials of such an ND-system are reviewed below.
- For *subsentential phrases*, replace taking their denotations (in arbitrary models, extensions) as meaning, by taking their **contributions** to the meanings (in our explication, derivability conditions) of sentences in which they occur. This adheres to Frege's *context principle* (Frege 1884), made more specific by the incorporation into a *type-logical grammar (TLG)* (see Moortgat 1997), the assumed underlying syntactic formalism. A detailed exposition of deriving meanings of subsentential phrases can be found in (Francez *et al.* 2010) for natural language, and in (Francez and Ben-Avi 2011) for logic.

According to the mainstream PTS programme, meaning is determined via a meaning-conferring natural-deduction proof system. An ND-system has two families of rules for each defined expression.

Introduction rules (I-rules): These are rules specifying the way a formula (sentence) having the defined expression as its main operator, the conclusion of the rule, can be deduced from other formulas, serving as premises of the rule. Such a deduction is the *most direct* way to deduce the conclusion.

Elimination rules (*E*-rules): These are rules specifying the way a consequence can be deduced from a formula (sentence) having the defined expression as its main operator, the major premise of the rule, and from some additional minor premises. Such a conclusion is the most direct conclusion of the major premise.

Both kind of rules can *discharge* assumptions, usually indicated with square brackets. Derivability of φ from a collection of assumptions (a *context*[3]) Γ is denoted $\Gamma \vdash \varphi$. Derivation $\mathscr{D}$ of φ from Γ is the usual recursively defined one. I use the Gentzen-Prawitz tree-like format for presenting derivations. One of its advantages over linear representations of ND-derivations, useful in the current discussion, is the convenience of representing the composition of derivations, needed for defining reductions.

An important requirement is that the ND-system should be *harmonious*, in that its rules have a certain balance between introduction and elimination, in order to qualify as conferring meaning. Harmony is delineated in more detail below.

A standard reference for ND-systems for logic is (Prawitz 1965). For ND-systems for an extensional ND-fragment, see (Francez and Dyckhoff 2010).

To explain the QDR-problem itself, consider the following example sentence from (from Stanley and Szabó 2000).[4]

$$\textsf{every bottle is empty} \tag{1.1}$$

The literal model-theoretic meaning of (1.1), involving quantification and predication, attributes the property of emptiness to every entity in a model falling under the extension of bottle.[5] This truth condition is usually expressed as the inclusion of the extension of bottle in the extension of empty, alluding to the generalized quantifiers theory. The general consent is, however, that in different circumstances, to be captured by contexts, the domain of quantification is not over the

[3] Not to be confused with a *DR*-context $c \in C$ affecting meaning variability, as defined below.

[4] All the examples of natural language expressions are depicted in the san-serif font, and are always mentioned, not used.

[5] As noted by Glanzberg (2006), it suffices to conduct this study in an extensional fragment of NL, as intentionality seems orthogonal to QDR-problem.

whole extension of bottle (all bottles in the universe); rather, it is over a restriction of this extension to one determined by a context, e.g., every bottle in a room where some party takes place in one context, or bottles in some chemistry laboratory in another context. Similarly,

$$\text{some bottle is empty} \tag{1.2}$$

is taken also to have contextually varying meaning, asserting that some bottle, determined by a given context, is empty, not that some bottle in the universe is empty.

A more radical approach, called *contextualism*, claims that *there is no quantification which is not contextually restricted*! Even apparently unrestricted quantification as expressed by everything or something are contextually restricted (see, for example, Glanzberg 2006).

Note that MTS in general adheres to a compositional sentential meaning assembly. The primary carriers of meaning are words, interpreted as having denotations in models (that can be rather complex), and semantic composition generates meanings for phrases until the meaning of a whole (affirmative) sentence is determined. According to this methodology, (some of) the word denotations are context-dependent, a dependence propagated to larger phrases as the interpretation process advances. I'll return to this issue in the sequel.

The *general* semantic problem faced in an attempt to model the variance of literal meaning with context has, according to Stanley and Szabó (2000), two facets.

Descriptive: Deriving the interpretation of some phrase relative to a context, given prior characterization of which features of a context have a bearing on the meaning of that phrase.

Fundamental: Specifying the above mentioned characterization, namely what it is about a context in virtue of which the derivation of the interpretation yields the correct meaning in that context. This specification involves some *explicit* definition of a context.

Thus, for (1.1), the descriptive meaning is the proper derivation of the restricted domain of quantification given a context, while the fundamental issue is what in the structure of a context determines the appropriate domain restriction.

In general, MTS has many difficulties in adequately solving the foundational aspect of contextual variance of truth conditions. A major contribution of the current paper is the provision of a solution, within the PTS programme, of the foundational problem.

In MTS, it is far from clear where to locate contexts with respect to a model. Stanley and Szabó (2000, p. 222), for example, admit that they avoid giving a formal characterization of the notion of a context. They just stipulate (for the QDR-problem) a certain marking *in the syntactic tree* (the logical form) that interfaces in a certain way with a context, and provide a description of the way this marking participates in meaning derivations (by intersecting the extension of the head noun with a set "pointed to" by the above mentioned marking). More specifically, Stanley and Szabó (2000) posit as the lexical entry of a noun, say man, (in the appropriate leaf of a syntactic tree) the following compound expression.

$$\langle man, f(i)\rangle \tag{1.3}$$

where *man* is the usual extension of man (in a model), i is an anchor for an object to be provided by a context, and f is an anchor to a function from objects to objects, also to be provided by context. The rule for computing the extension of man in a given context c is the following (in a slightly modified notation).

$$[\![\langle man, f(i)\rangle]\!]_c \overset{\text{df.}}{=} [\![man]\!] \cap \{x \mid x \in c[\![f]\!](c[\![i]\!])\} \tag{1.4}$$

For an argument for a different location (in the syntactic tree) of that marking (and for a rebuttal of the rejection of this location by Stanley and Szabó 2000), see (Pelletier 2003). There are also views locating this marker on the determiner node, (e.g., Westerståhl 1985). Note that in all those approaches, there is no constraint at all imposed on the set $\{x \mid x \in c[\![f]\!](c[\![i]\!])\}$. In particular, as is traditional in generalized quantifier theory, this set need not be the extension (in the model at hand) of any NL phrase.

I would like to claim that this degree of freedom regarding the contextual restriction set is a drawback of all the above approaches to QDR. In general, a context can be seen either as external to the interpreted sentence (e,g., a context of utterance), or explicitly contributed by some phrase in the sentence itself. For example, (1.1) can be seen as uttered in the context of bottles on some table; however, the location of the bottles can be explicitly given in the sentence itself, say by means of a preposition phrase, as in

every bottle on the table is empty (1.5)

Furthermore, if the intended context is such that the salient bottles are bottles of whisky, then this again can be given by an additional explicit modification of the noun, as in

every whisky bottle on the table is empty (1.6)

I would like to posit the following *context incorporation* principle as a characterization of contextually varying meaning (as far as QDR is concerned). I see this principle as originating from the semantic concept of 'meaning' (as far as it relates to contextual meaning variation), and not from any empirical fact about this variation. One certainly can conceive of contexts not having any linguistic expression. As I see it, while such contexts can contribute to other dimensions of language use, alluding to them is *not* part of meaning.

The context incorporation principle (CIP): For every quantified sentence S with a meaning depending on a context c, there exists a (not necessarily unique) sentence S', s.t.

$$[\![S]\!]_c = [\![S']\!] \tag{1.7}$$

In other words, the effect of a *given* external context c in terms of QDR in S is always expressible by S', the meaning of which is independent of c (all in the same language, or fragment thereof). Clearly, (CIP), while being allowed by (1.4), is not *enforced* by (1.4).

It is important to realise what *is not* the semantic problem discussed here, namely *the determination of which is the right context for any given token of a contextually dependent meaning of a sentence*. The latter issue is always determined by extra-linguistic means, independently of whether MTS or PTS are employed as the theory of meaning. Rather, the issue is how to handle contextual meaning variation once a context has been determined. Thus, if the intended context for the above example is bottles of whisky, then an explicit assumption to this effect has to be added to the given context. Once the *whole* intended context has been incorporated, the resulting sentence should be read as context independent.

Finally, the *consequences* that can be drawn from the contextually varying meaning of an (affirmative) sentence, namely (affirmative) sentences *entailed* by a sentence with contextually varying meaning, which themselves have meanings varying with context, are hardly ever considered in MTS-based discussions. I will relate to them in the proposed PTS via E-rules in the meaning-conferring ND-system.

Note that I adopt here the view expressed in (Stanley and Szabó 2000) that contextual variance of meaning is a *semantic* phenomenon, and not a syntactic (ellipsis) or pragmatic (agent related) one. I would like to stress that I am investigating *what (affirmative) sentences mean, and how this meaning varies with context*, and *not* with what an agent means by asserting a sentence in a given context; the latter, involving intentions, plans etc., I do see as pragmatic. Thus, I exclude from consideration examples such as the following (from Stanley and Szabó 2000)

$$\textsf{Fred is a good friend} \tag{1.8}$$

uttered by a speaker in some circumstances to express that Fred is, actually, a terrible friend. I do not take this interpretation of (1.8) as a *meaning* of (1.8) in any sense of 'meaning' that semantics is concerned with.

Why adhere to CIP?

- One can see the semantic view of the QDR-phenomenon alluded above as a (partial) justification of CIP, that relates to linguistically expressible contexts. In a performative, agent related, use of a sentence with a contextually varying meaning, it is conceivable that other kinds of contextual information, not language oriented, may have an effect. For example, complex visual information in a common ground of speaker and hearer. This is certainly true for contextual resolution of deictic elements in a sentence. This would pertain to context dependence of meaning that fits a more traditional view of it, as pragmatic, not semantic.
- While I am concerned here with meanings of single (affirmative) sentences, there is clearly much semantic interest in *dialogs*, or *discourses*, which are multi-sentential linguistic entities. Adhering to CIP is compatible with identifying context with the contents of sentences previously asserted by other participants in a dialog, or preceding sentences in a discourse. From my proof-theoretic point

of view, the PTS for such multi-sentential linguistic constructs is, at best, in its infancy. Principles like CIP may encourage further proof-theoretic investigations of such constructs.

The paper contains also a certain proof-theoretic innovation in the concept of a parametric family of introduction rules (in a natural-deduction system), which is not directly connected to the NL set-up aimed at in the paper.

In (Francez and Wieckowski 2014), a similar approach to contextual meaning variation is applied to *contextual definiteness*, as in

the bottle is empty (1.9)

where the usual existence and uniqueness, traditionally associated with definiteness, is restricted to given contexts.

In the rest of this paper, I provide a PTS for the QDR-problem, relating both to its fundamental facet as well as to its descriptive facet, by providing meaning-conferring ND-systems. I start in Section 2 with casting the solution in a *logic* setting, its familiarity facilitating a clearer explication of the proof-theoretic technique involved. Then, I consider a PTS for the incorporation of the QDR-problem in an extensional fragment of English, for which a PTS is provided in (Francez and Dyckhoff 2010). The paper ends with some conclusions.

2 LOGIC WITH CONTEXTUAL DOMAIN RESTRICTION

In this section, I present a version of first-order logic (FOL) in which quantifiers are interpreted in a contextually dependent way. While there is not much interest in such a logic per se, it serves as a vehicle for a clear presentation of the ideas underlying the application of the approach to natural language. It also provides a natural host for the novel proof-theoretic concept of a *parameterized family of I-rules* in the intended natural-deduction meaning-conferring proof system.

2.1 *First-order logic with contextual domain restriction*

I assume the usual object language for FOL, with the usual definition of free/bound variables. For simplicity, a language without constant or function symbols is considered.

Definition 2.1 (DR-context) A *DR-context* (domain restricting context) c is a finite collection Γ_c of open formulas with one free variable only; $\Gamma_{c,x}$ is the sub-collection of Γ_c with x as its free variable. Let C be the collection of all DR-contexts.

This definition of a DR-context is certainly not the most general one for a context affecting sentential meanings, but it is intended to capture contexts as providing restriction on quantifiers, for which purpose this definition suffices. Let $\wedge_{\Gamma_{c,x}}$ be the conjunction of all open formulas in $\Gamma_{c,x}$ (that have x free). I use $\Gamma_{c,x}$ and $\wedge_{\Gamma_{c,x}}$ interchangeably. I use $\Gamma_{c,x}(y)$ or $\wedge_{\Gamma_{c,x}}(y)$ to indicate the application of the condition on x to another variable, y, resulting in a substitution of y for free occurrences of x.

The main idea, to be captured by the rules below, is that a DR-context provides an assumption, *dischargeable* in the case of universal quantification, restricting the free variable in the premise of the I-rule of a quantifier. Furthermore, this discharge keeps its contents excorporated from the formula (recording the context c generating it in '$\vdash_c$').

First, recall the standard I/E rules for the universal and existential quantifiers in an ND-system for FOL.[6]

$$\frac{\Gamma \vdash \varphi(x)}{\Gamma \vdash \forall x.\varphi(x)}\ (\forall I),\ x \notin \mathit{free}(\Gamma) \qquad \frac{\Gamma \vdash \forall x.\varphi(x)}{\Gamma \vdash \varphi(y)}\ (\forall E) \tag{2.10}$$

$$\begin{gathered}\frac{\Gamma \vdash \varphi(y)}{\Gamma \vdash \exists x.\varphi(x)}\ (\exists I)\\ \frac{\Gamma \vdash \exists x.\varphi(x) \quad \Gamma, [\varphi(y)]_i \vdash \chi}{\Gamma \vdash \chi}\ (\exists E^i),\ y \notin \mathit{free}(\Gamma, \chi)\end{gathered} \tag{2.11}$$

where $\varphi(y)$ is the result of substituting y for all free occurrences of x in $\varphi(x)$. I now introduce a revised ND-system, in which deducibility is indicated as '$\vdash_c$' (in contrast to '$\vdash$' indicating the deducibility in the standard system).

Restricting the universal quantifier: Recall that the intuition behind the usual $(\forall I)$-rule is that since x does not occur free in Γ, it can be seen as standing for an *arbitrary* value, unrestricted in any way by Γ,

[6] I assume familiarity with standard I/E-rules for the propositional operators, like conjunction '$\wedge$' and implication '$\rightarrow$'; see (Prawitz 1965) for a standard presentation.

hence supporting the universal generalization embodied in the $(\forall I)$-rule. The idea behind the I-rule below is to restrict the generalization to those values of x satisfying the contextual restriction imposed by $\Gamma_{c,x}(x)$ for a given DR-context c. Thereby, the *same* formula $\forall x.\varphi(x)$ is read differently in different DR-contexts. This is achieved by using $\Gamma_{c,x}(x)$ as a discharged assumption in the premise of the rule.

$$\begin{gathered}\frac{\Gamma, [\Gamma_{c,x}(x)]_i \vdash_c \varphi(x)}{\Gamma \vdash_c \forall x.\varphi(x)} (\forall I^i_C),\ x \notin free(\Gamma)\\ \frac{\Gamma \vdash_c \forall x.\varphi(x) \quad \Gamma \vdash_c \wedge_{\Gamma_{c,x}}(y)}{\Gamma \vdash_c \varphi(y)} (\forall E_C)\end{gathered} \tag{2.12}$$

Here $\forall I_C$ is a family of I-rules indexed by DR-contexts. Every application of this rule is always by appealing to some given DR-context $c \in C$. In the interesting cases, $\Gamma_{c,x} \neq \emptyset$ will hold, though there might be vacuous DR-contexts not affecting the meaning of a universal sentence. Similarly, $(\forall E_C)$ is a family of E-rules indexed by DR-contexts. The conclusion drawn from $\forall x.\varphi(x)$ deduced relative to a DR-context c is read as $\wedge_{\Gamma_{c,x}}(y) \rightarrow \varphi(y)$, namely that y satisfies both $\varphi(x)$ and the contextual restriction $\wedge_{\Gamma_{c,x}}(x)$.

Restricting the existential quantifier: As for existential quantification, the contextual rules are presented below.

$$\begin{gathered}\frac{\Gamma \vdash_c \varphi(y) \quad \Gamma \vdash_c \wedge_{\Gamma_{c,x}}(y)}{\Gamma \vdash_c \exists x.\varphi(x)} (\exists I_C)\\ \frac{\Gamma \vdash_c \exists x.\varphi(x) \quad \Gamma, [\varphi(y)]_i, [\wedge_{\Gamma_{c,x}}(y)]_j \vdash_c \chi}{\Gamma \vdash_c \chi} (\exists E^{i,j}),\ y \notin free(\Gamma, \chi)\end{gathered} \tag{2.13}$$

The I-rule requires that for some y that satisfies the restrictions imposed by $\Gamma_{c,x}$, $\varphi(y)$ is derived, in order to deduce that the contextually restricted (by c) existential conclusion be derived. Recall that, like in the standard $(\exists I)$-rule, y may, (and in general, will) appear free in Γ. So, the rule forces y to also fall under the restriction imposed by c. The E-rule, like the standard $(\exists E)$-rule, allows the derivation of an arbitrary conclusion χ, under the assumption that φ and the contextual restriction themselves derive χ (for a fresh y).

Remark: From the above rules, it is evident that (CIP) holds for FOL with QDR. This is true since $\Gamma_{c,x}$ (and consequently, $\bigwedge_{\Gamma_{c,x}}$) consist only of formulas in the language.

Theorem 2.1 (context incorporation)

1. $\Gamma \vdash_c \forall x.\varphi(x)$ iff $\Gamma \vdash \forall x. \bigwedge_{\Gamma_{c,x}}(x) \to \varphi(x)$.
2. $\Gamma \vdash_c \exists x.\varphi(x)$ iff $\Gamma \vdash \exists x. \bigwedge_{\Gamma_{c,x}}(x) \wedge \varphi(x)$.

Proof:

1. (a) Assume $\Gamma \vdash_c \forall x.\varphi(x)$ is derived by means of $(\forall I_C)$. By an inductive argument, the premise of $(\forall I_C)$, namely $\Gamma, [\Gamma_{c,x}(x)]_i \vdash_c \varphi(x)$ (with $x \notin \mathit{free}(\Gamma)$), implies that $\Gamma, [\Gamma_{c,x}(x)]_i \vdash \varphi(x)$. Therefore, by using $(\to I_i)$, $\Gamma \vdash \bigwedge_{\Gamma_{c,x}(x)} \to \varphi(x)$, and by applying $(\forall I)$ (since $x \notin \mathit{free}(\Gamma)$), we get $\Gamma \vdash \forall x.\bigwedge_{\Gamma_{c,x}(x)} \to \varphi(x)$.
   (b) Conversely, suppose $\Gamma \vdash \forall x.\bigwedge_{\Gamma_{c,x}(x)} \to \varphi(x)$ is derived via $(\forall I)$ with the premise $\Gamma \vdash \bigwedge_{\Gamma_{c,x}(x)} \to \varphi(x)$, where $x \notin \mathit{free}(\Gamma)$. Thus, also $\Gamma, [\Gamma_{c,x}(x)]_i \vdash \varphi(x)$ (due to $(\to I)$). By an application of $(\forall I_C^i)$ the result follows.
2. The argument for existential quantification is similar and omitted.

Here are some examples for the more interesting direction.

1. In the DR-context c, $\forall x.\varphi(x)$ is read as $\forall x. \bigwedge_{\Gamma_{c,x}}(x) \to \varphi(x)$. When $\varphi(x)$ is itself an implication, say $\alpha(x) \to \beta(x)$, then the result is equivalent to conjoining the antecedent with the contextual restriction, $\forall x.\alpha(x) \wedge \bigwedge_{\Gamma_{c,x}}(x) \to \beta(x)$.
2. Similarly, in the DR-context c, $\exists x.\varphi(x)$ is read as $\exists x.\bigwedge_{\Gamma_{c,x}}(x) \wedge \varphi(x)$.

Example 2.1 Suppose (1.1) is regimented by the FOL-formula $\forall x.B(x) \to E(x)$ (with $B(x)$ interpreted as x is a bottle and $E(x)$ as x is empty). Let c_{room} be a DR-context of some room, with $\Gamma_{c_{room},x} = \{R(x)\}$ (with $R(x)$ interpreted as x is in the room). Then,

$$\frac{\Gamma, [R(x)]_i \vdash_{c_{room}} B(x) \to E(x)}{\Gamma \vdash_{c_{room}} \forall x.B(x) \to E(x)} \; (\forall I_C^i), \; x \notin \mathit{free}(\Gamma)$$

allows the derivation of a reading of (1.1) as $\forall x.R(x) \to (B(x) \to E(x))$, equivalent to $\forall x.B(x) \wedge R(x) \to E(x)$; that is, every bottle in the room is empty. This can be seen as incorporating the DR-context into the sentence. Note that the contextually derived universally quantified sentence does not carry its contextual meaning “on its nose”. To obtain the required reading, one has to know the DR-context in which the sentence was derived (c_{room} in this example), and consult $\Gamma_{c_{room},x}$ to obtain this reading. Similarly,

$$\dfrac{\Gamma \vdash_{c_{room}} \forall x.B(x) \rightarrow E(x) \quad \Gamma \vdash_{c_{room}} R(y)}{\Gamma \vdash_{c_{room}} B(y) \rightarrow E(y)}\;(\forall E_C)$$

allows drawing from (1.1) derived in the DR-context c_{room} the conclusion $R(y) \rightarrow (B(y) \rightarrow E(y))$, equivalent to $B(y) \wedge R(y) \rightarrow E(y)$; namely a reading corresponding to if y is a bottle in the room then y is empty, a correct reading of the conclusion in the context c_{room}.

Example 2.2 Following is another example, establishing

$$\forall x.W(x) \wedge I(x) \rightarrow S(x), \forall y.W(y) \wedge S(y) \rightarrow B(y) \\ \vdash_c \forall z.W(z) \rightarrow B(z) \quad (2.14)$$

in a DR-context c with $\Gamma_{c,z} = I(z)$. I'll return to this example below.

$$\dfrac{\dfrac{\dfrac{\dfrac{[W(z)]_1 \quad \dfrac{\dfrac{[W(z)]_1 \quad [I(z)]_2}{W(z) \wedge I(z)}\,(\wedge I) \quad \dfrac{\forall x.W(x) \wedge I(x) \rightarrow S(x)}{W(z) \wedge I(z) \rightarrow S(z)}\,(\forall E)}{S(z)}\,(\rightarrow E)}{W(z) \wedge S(z)}\,(\wedge I) \quad \dfrac{\forall y.W(y) \wedge S(y) \rightarrow B(y)}{W(z) \wedge S(z) \rightarrow B(z)}\,(\forall E)}{B(z)}\,(\rightarrow E)}{W(z) \rightarrow B(z)}\,(\rightarrow I^1)}{\forall z.W(z) \rightarrow B(z)}\,(\forall I_C^2) \quad (2.15)$$

Example 2.3 The next example is of two independent QDRs by a context. It shows why the premises of the I_C-rules themselves have to use '$\vdash_c$', and not just '$\vdash$'.

$$\forall x \forall y.M(x) \wedge Y(x) \wedge W(y) \wedge S(y) \rightarrow L(x,y), \\ \forall z.W(z) \wedge I(z) \rightarrow S(z) \\ \vdash_c \forall x \forall y.M(x) \wedge W(y) \rightarrow L(x,y) \quad (2.16)$$

where $\Gamma_{c,x} = Y(x)$, $\Gamma_{c,y} = I(y)$. Let **I, II** abbreviate, respectively, the two premises. I treat '$\wedge$' as having arbitrary arity.

$$\cfrac{\cfrac{\cfrac{\cfrac{\cfrac{\mathbf{I}}{M(x)\wedge Y(x)\wedge W(y)\wedge S(y)\rightarrow L(x,y)}\,(\forall E)\times 2 \quad \cfrac{\cfrac{[M(x)\wedge W(y)]_3}{M(x)}\,(\wedge E) \quad [Y(x)]_1 \quad \cfrac{[M(x)\wedge W(y)]_3}{W(y)}\,(\wedge E) \quad \cfrac{\cfrac{\mathbf{II}}{W(y)\wedge I(y)\rightarrow S(y)}\,(\forall E) \quad \cfrac{\cfrac{[M(x)\wedge W(y)]_3}{W(y)}\,(\wedge E) \quad [I(y)]_2}{W(y)\wedge I(y)}\,(\wedge I)}{S(y)}\,(\rightarrow E)}{M(x)\wedge Y(x)\wedge W(y)\wedge S(y)}\,(\wedge I)}{L(x,y)}\,(\rightarrow E)}{M(x)\wedge W(y)\rightarrow L(x,y)}\,(\rightarrow I^3)}{\forall y.M(x)\wedge W(y)\rightarrow L(x,y)}\,(\forall I^2_C)}{\forall x\forall y.M(x)\wedge W(y)\rightarrow L(x,y)}\,(\forall I^1_C) \tag{2.17}$$

The following proposition expresses a property of the QDR-rules that will be useful below. It says that it does not matter which variable is used to express the contextual restriction, as long as it is amenable to universal generalization.

Proposition 2.1 If $\Gamma, [\Gamma_{c,x}(x)]_i \vdash_c \varphi(x)$ and $y \notin \mathit{free}(\Gamma)$, then also $\Gamma, [\Gamma_{c,x}(y)]_i \vdash_c \varphi(y)$.

Next, consider the definition of the (reified) contextually varying sentential meanings, following the ideas in (Francez 2014c).

Definition 2.2 (**canonical derivation**) A derivation $\mathscr{D}$ for $\Gamma \vdash \psi$ is *canonical* iff it satisfies one of the following two conditions.

- The last rule applied in $\mathscr{D}$ is an I-rule (for the main operator of ψ).
- The last rule applied in $\mathscr{D}$ is an assumption-discharging E-rule, the major premise of which is some φ in Γ, and its encompassed sub-derivations $\mathscr{D}_1, \cdots, \mathscr{D}_n$ are all canonical derivations of ψ.

Canonical derivations constitute the most *direct* derivations of their conclusion (though not necessarily always the shortest), and are viewed by PTS to underlie and determine meaning. Let $[\![\varphi]\!]^c_\Gamma$ denote the (possibly empty) collection of all canonical derivations of φ from Γ.[7]

Definition 2.3 (**reified meanings**) The *(reified) meaning* of φ is given by

$$[\![\varphi]\!] \stackrel{\text{df.}}{=} \lambda\Gamma.[\![\varphi]\!]^c_\Gamma \tag{2.18}$$

To realize the role of canonicity in the definition of reified proof-theoretic meanings, consider the following example derivation in

[7] The superscript 'c' here relates to canonicity, and should not be confused with a DR-context, the latter indicated by a subscript c.

propositional logic.

$$\frac{\alpha \quad (\alpha \rightarrow (\varphi \wedge \psi))}{\varphi \wedge \psi}\ (\rightarrow E) \tag{2.19}$$

This is a derivation of a conjunction – but not a canonical one, as it does not end with an application of $(\wedge I)$. Thus, the conjunction here was *not* derived according to its meaning! As far as this derivation is concerned, it could mean anything, for example, disjunction. On the other hand, the following example derivation, being canonical, *is* according to the conjunction's meaning.

$$\frac{\dfrac{\alpha \quad \alpha \rightarrow \varphi}{\varphi}\ (\rightarrow E) \quad \dfrac{\beta \quad \beta \rightarrow \psi}{\psi}\ (\rightarrow E)}{\varphi \wedge \psi}\ (\wedge I) \tag{2.20}$$

Similar examples can be found in natural language.

We can now see the difference between ordinary meanings and their contextually varying counterpart. For the context-independent meaning of $\forall x.\varphi(x)$, all the canonical derivations end with an application of the *same* $(\forall I)$-rule, while for the meaning of $\forall x.\varphi(x)$ in a DR-context c, all canonical derivations end with an application of $(\forall I_C)$, varying with c.

As was already observed in (Francez 2014a), this reified meaning is very fine-grained,[8] and a certain relaxation of it is found useful. Note that the CIP requires (strict) sameness of meaning between a contextually restricted quantified sentence and its context incorporated counterpart. However, while the relationship of canonical derivations of both are very similar – the former ending with application of $(\forall I_C)$ (in the universal case) whenever the latter ends with $(\rightarrow I)$ immediately followed by $(\forall I)$, they are strictly not the same! We can obtain a natural coarsening fitting also the current needs (for the CIP), still fine enough as to not identify the meanings of logically equivalent sentences as done in MTS, by introducing *grounds (for assertion)* for sentences (see Francez and Dyckhoff 2010 and Francez 2014c for a discussion of the role of those grounds in the PTS programme).

[8] For example, it is shown in (Francez 2014a) that $[\![\varphi \wedge (\psi \wedge \chi)]\!] \neq [\![(\varphi \wedge \psi) \wedge \chi]\!]$.

Definition 2.4 (grounds for assertion) The *grounds for assertion* of φ, denoted by $G[\![\varphi]\!]$, are given by

$$G[\![\varphi]\!] \overset{\text{df.}}{=} \{\Gamma \mid \Gamma \vdash^c \varphi\} \tag{2.21}$$

In other words, any Γ from which there is a canonical derivation of φ serves as a ground for asserting φ.

I now introduce an equivalence relation on meaning based on *sameness of grounds (for assertion)*, that captures the CIP requirement in a natural way.

Definition 2.5 (grounds equivalence)

$$\varphi \equiv_G \psi \text{ iff } G[\![\varphi]\!] = G[\![\psi]\!] \tag{2.22}$$

Obviously, '$\equiv_G$' is an equivalence relation on meanings. An easy inspection of the proof of the context incorporation theorem shows that the meanings of the context-incorporated counterparts of contextually restricted quantified sentences are grounds equivalent.

2.2 *Harmony of the contextual domain restriction rules*

Prior's famous attack on the PTS-programme in (Prior 1960) produced a connective with an I-rule of disjunction and an E-rule of conjunction, that trivialized '$\vdash$' so that $\varphi \vdash \psi$ for *every* φ and ψ. As became evident since that attack, not every set of I/E-rules may qualify as conferring meaning. One of the prevailing criteria for an ND-system to qualify as conferring meaning is that of *harmony*, advocated by Dummett, Prawitz, Tennant and many others, requiring a *balance* between the I-rules and E-rules of every connective, in that neither group is either too weak or too strong w.r.t. the other group. Clearly, Prior's connective fails this condition. Two main ways to capture the informal notion of harmony were proposed in the literature.

Intrinsic harmony: According to this view of harmony, there is a requirement that every *maximal formula* φ in a derivation, one that is a conclusion of an I-rule and the major premise of an E-rule (both of the main operator of φ), be eliminable, producing an equivalent derivation (with the same assumptions and same conclusion). The process of eliminating such a maximal formula is known as *(proof) reduction*, and underlies Prawitz's *normalization* procedure (Prawitz 1965).[9] Reductions show that nothing is gained by introducing and immediately eliminating. The gain here does not refer to efficiency (mostly lengths) of derivations, but to the ability to draw additional conclusion.[10] In a balanced system, any conclusion drawn by means of a maximal formula can be drawn without it, as shown by the reduction. Failing this condition shows that the I-rule is too strong (or the E-rule too weak). The second facet of the balance between I/E-rules is that of *stability*, excluding a situation in which the E-rules are *too weak* w.r.t. the I-rules. I will ignore this issue here.

Harmony in form: Under this view of harmony, the E-rules are required to have a *specific form*, known as *general elimination* (*GE*), allowing the derivation of an *arbitrary conclusion* using the premises of the I-rules as discharged assumptions. The standard rules $(\vee E)$ and $(\exists E)$ are of this form. *GE*-rules emerged independently of harmony, allowing a better correspondence between normal ND-derivations and *CUT*-free derivations in sequent-calculi (see, for example, Schroeder-Heister 1984; von Plato 2000, 2001). In (Francez and Dyckhoff 2012) a general procedure[11] is presented for deriving *harmoniously induced GE*-rules from *given* I-rules, ensuring the availability of the reductions required by intrinsic harmony.

Below, I show the reductions for the rules for '$\vdash_c$'.

Universal contextually restricted quantification:

$$\dfrac{\dfrac{\Gamma, [\Gamma_{c,x}(x)]_i \vdash_c \varphi(x)}{\Gamma \vdash_c \forall x.\varphi(x)}\,(\forall I_C^i) \qquad \Gamma \vdash_c \wedge_{\Gamma_{c,x}}(y)}{\Gamma \vdash_c \varphi(y)}\,(\forall E_C)$$

$$\rightsquigarrow_r \quad \Gamma[x := y], [\Gamma_{c,x}(y)]_i \vdash_c \varphi(y) \tag{2.23}$$

[9] Note that the presence of a reduction is less demanding than normalisation. The latter requires the finiteness of reduction sequences.

[10] Often, efficient derivation are not according to the meaning determined by I-rules. For example, if one first proves $\forall x.\varphi(x)$, and then derives (via $\forall E$) $\varphi(a), \varphi(b)$ etc., the derivations of the latter are shorter, but not according to meaning.

[11] Recently, some restrictions on the domain of applicability of this procedure have been realized, but they do not affect the current set-up.

Note that since $x \notin \mathit{free}(\Gamma)$, $\Gamma[x := y] = \Gamma$. The result follows by Proposition 2.1. A clearer depiction of the reduction uses $\mathscr{D}$s.

$$\dfrac{\dfrac{\begin{matrix}[\wedge_{\Gamma_{c,x}}(x)]_i\\ \mathscr{D}\end{matrix}}{\forall x.\varphi(x)}\,(\forall I^i_C) \qquad \begin{matrix}\mathscr{D}'\\ \wedge_{\Gamma_{c,x}}(y)\end{matrix}}{\varphi(y)}\,(\forall E_C) \quad \leadsto_r \quad \begin{matrix}\mathscr{D}'\\ \mathscr{D}[\wedge_{\Gamma_{c,x}}(y) := \wedge_{\Gamma_{c,x}}(y)]\\ \varphi(y)\end{matrix} \tag{2.24}$$

where the substitution $\wedge_{\Gamma_{c,x}}(y) := \overset{\mathscr{D}'}{\wedge_{\Gamma_{c,x}}(y)}$ is the usual way composition of derivations is obtained, by replacing an assumption $\wedge_{\Gamma_c}(y)$ (a leaf in $\mathscr{D}$) with its given derivation in the second premise of $(\forall E_C)$.
The harmoniously induced $(\forall GE_C)$ is given below.

$$\dfrac{\Gamma \vdash_c \forall x.\varphi(x) \quad \Gamma \vdash_c \wedge_{\Gamma_{c,x}}(y) \quad \Gamma, [\varphi(y)]_i \vdash_c \chi}{\Gamma \vdash_c \chi}\,(\forall GE^i_C), \quad y \text{ fresh} \tag{2.25}$$

Existential contextually restricted quantification: The reduction is the following.

$$\dfrac{\dfrac{\begin{matrix}\mathscr{D}_1\\ \varphi(y)\end{matrix} \quad \begin{matrix}\mathscr{D}_2\\ \wedge_{\Gamma_{c,x}}(y)\end{matrix}}{\exists x.\varphi(x)}\,(\exists I_C) \qquad \begin{matrix}[\varphi(z)]_i, [\wedge_{\Gamma_{c,x}}(z)]_j\\ \mathscr{D}\\ \chi\end{matrix}}{\chi}\,(\exists E^{i,j}_C)$$

$$\leadsto_r \quad \begin{matrix} \mathscr{D}[\varphi(z) := \begin{matrix}\mathscr{D}_1[y := z]\\ \varphi(z)\end{matrix}, \wedge_{\Gamma_{c,x}}(z) := \begin{matrix}\mathscr{D}_2[y := z]\\ \wedge_{\Gamma_{c,x}}(z)\end{matrix}]\\ \chi\end{matrix} \tag{2.26}$$

The $(\exists E_C)$-rule is in the *GE*-form to start with, thus harmonious in form too.

2.3 *Quantifier domain restriction in restricted quantification*

In order to make the subsequent presentation of QDR in NL more comprehensible, I exemplify the proof-theoretic approach by applying it first to a fragment FOL_{rq} of FOL that comes closer to the NL-fragment to be considered. The fragment is known as having *restricted quantification* (not to be confused with contextually restricted quantification, which is added on top of this). Quantified formulas in this fragment have the following form:

$$\forall x.\varphi(x) \to \psi(x), \;\; \exists x.\varphi(x) \wedge \psi(x) \tag{2.27}$$

$$\frac{}{\Gamma,\xi \vdash \xi}\ (Ax)$$

$$\frac{\Gamma,[\varphi(y)]_i \vdash \psi(y)}{\Gamma \vdash \forall x.\varphi(x) \to \psi(x)}\ (\forall I^i) \qquad \frac{\Gamma \vdash \varphi(y) \quad \Gamma \vdash \psi(y)}{\Gamma \vdash \exists x.\varphi(x) \wedge \psi(x)}\ (\exists I)$$

y fresh for Γ in $(\forall I)$.

$$\frac{\Gamma \vdash \forall x.\varphi(x) \to \psi(x) \quad \Gamma \vdash \varphi(y)}{\Gamma \vdash \psi(y)}\ (\forall E) \qquad \frac{\Gamma \vdash \exists x.\varphi(x) \wedge \psi(x) \quad \Gamma,[\varphi(y)]_i,[\psi(y)]_j \vdash \xi}{\Gamma \vdash \xi}\ (\exists E^{i,j})$$

y fresh for Γ, ξ in $(\exists E)$.

Figure 1: A natural-deduction proof system N_{rq} for restricted quantification

The universal quantification can be read as 'everything which is φ is ψ', and the existential quantification can be read as 'there exists something which is φ that is ψ'. That is, quantification is restricted to entities satisfying φ, to be called the *restrictor*. A more transparent syntax, closer to the natural language expression of quantification, would be

$$\forall x : \varphi(x).\psi(x), \quad \exists x : \varphi(x).\psi(x) \tag{2.28}$$

The expression of (2.28) as (2.27) is known as Frege's translation, that has drawn criticism as a way to capture natural language quantification. For example, see (Ben-Yami 2006) and (Francez 2014b) for such a criticism. As I show in the next section, FOL$_{rq}$-quantification reflects more directly natural language quantification.

The proof system is presented in Figure 1. I use $\Gamma \vdash \varphi$ in this subsection to indicate derivability of φ from Γ (in N_{rq}). A *GE*-rule for the universal quantifier, exhibiting harmony in form, is

$$\frac{\Gamma \vdash \forall x.\varphi(x) \to \psi(x) \quad \Gamma \vdash \varphi(y) \quad \Gamma,[\psi(y)]_i \vdash \xi}{\Gamma \vdash \xi}\ (\forall GE^i) \tag{2.29}$$

y fresh

Next, I consider QDR in FOL$_{rq}$. The observation is that the restrictor can be interpreted differently in different DR-contexts. Thus, the natural regimentation of (1.1) (cf. Example (2.1)) would again be

$$\forall x.B(x) \to E(x) \tag{2.30}$$

where '$B(x)$' expresses x is a bottle and '$E(x)$' expresses x is empty. Here, the restrictor $B(x)$ can have a contextually varying interpretation.

The idea for the proof-theoretic representation of contextual meaning variation is as before, where for universal quantification a DR-context c provides a contextual discharged assumption. The generated contextual restriction *strengthens* the restriction already present in the formula. The rules are shown below.[12]

$$\begin{array}{c} \dfrac{\Gamma, [\varphi(x)]_j, [\Gamma_{c,x}(x)]_i \vdash_c \psi(x)}{\Gamma \vdash_c \forall x.\varphi(x) \rightarrow \psi(x)}\ (\forall I_C^{i,j}),\quad x \notin \mathit{free}(\Gamma) \\[2ex] \dfrac{\Gamma \vdash_c \forall x.\varphi(x) \rightarrow \psi(x) \quad \Gamma \vdash_c \varphi(y) \wedge \wedge_{\Gamma_{c,x}}(y)}{\Gamma \vdash_c \psi(y)}\ (\forall E_C) \end{array} \tag{2.31}$$

$$\begin{array}{c} \dfrac{\Gamma \vdash_c \varphi(y) \quad \Gamma \vdash_c \Gamma_{c,x}(y) \quad \Gamma \vdash \psi(y)}{\Gamma \vdash_c \exists x.\varphi(x) \wedge \psi(x)}\ (\exists I_C^i) \\[2ex] \dfrac{\Gamma \vdash_c \exists x.\varphi(x) \wedge \psi(x) \quad \Gamma, [(\varphi \wedge \wedge_{\Gamma_{c,x}} \wedge \psi)(y)]_i \vdash_c \chi}{\Gamma \vdash_c \chi}\ (\exists E^i) \end{array} \tag{2.32}$$

where $y \notin \mathit{free}(\Gamma, \chi)$ in $(\exists E)$.

The same considerations as those for FOL show that (CIP) holds also for FOL_{rq}.

3 PROOF-THEORETIC SEMANTICS FOR QUANTIFIER DOMAIN RESTRICTION IN A FRAGMENT OF ENGLISH

In this section, I present a PTS for QDR in its more natural setting, within an extensional fragment of English. A PTS for such a fragment (without considering QDR) is provided in (Francez and Dyckhoff 2010).

For self-containment of the paper, this semantics is reviewed below.

3.1 *Review of the proof-theoretic semantics for sentences*

I present the fragment and its associated proof system in two stages. First, a core fragment is presented, extended in a second stage with relative clauses and intersective adjectives.

[12] The notation $\varphi(y) \wedge \wedge_{\Gamma_{c,x}}(y)$ means the conjunction of $\varphi(y)$ with the conjunction of the context formulas in $\Gamma_{c,x}$ applied to y.

3.1.1 The core fragment and proof system

The core fragment E_0^+ of English consists of sentences headed by (extensional) intransitive and transitive verbs, and determiner phrases (*dp*) with a (singular, count) noun and a determiner. In addition, there is the copula. This is a typical fragment of many NLs, syntactically focusing on *subcategorization*, and semantically focusing on *predication* and *quantification*. Some typical sentences are listed below.

$$\textsf{every/some girl smiles/is a student/loves some boy} \qquad (3.33)$$

I omit here *proper names* that do appear in the detailed presentation of sentential meanings (Francez and Dyckhoff 2010). Note the absence of *negative determiners* like no (hence the superscript '+'), which are treated in (Francez and Ben-Avi 2014), involving technicalities orthogonal to QDR. Expressions such as every girl, some boy are *dp*s.

The *PTS* is based on a core dedicated, meaning-conferring natural-deduction proof system N_0^+ with I/E-rules presented in Figure 2. The proof system is formulated over the language L_0^+, slightly extending E_0^+ and *disambiguating* ambiguous E_0^+ sentences. Meta-variables X schematize nouns, P over intransitive verbs and R over transitive verbs. Meta-variable S ranges over sentences, and boldface lower-case **j**, **k**, etc., range over $\mathscr{P}$, a denumerable set of *(individual) parameters*, artefacts of the proof system (not used to make assertions). Syntactically, a parameter in L_0^+ is also regarded as a *dp*. If a parameter occurs in S in some position, S is a *pseudo-sentence*, and if *all dps* in S are parameters, the pseudo-sentence S is *ground*. The ground pseudo-sentences play the role of atomic sentences, and their meaning is assumed *given*, externally to the ND proof system. The latter defines sentential meanings of non-ground pseudo-sentences (and, in particular, E_0^+-sentences), *relative* to the given meanings of ground pseudo-sentences.

In contrast to logic, where the introduced operator by an I-rule is always the (unique) *main operator*, in E_0^+ sentences there is no such main operator: every position that can be filled with a *dp* is a *locus of introduction* (of the quantifier corresponding to the determiner of the introduced *dp*). This is a major source of *ambiguity* in E_0^+, known as quantifier-scope ambiguity. The way ambiguity is treated is recapitulated briefly below. For any *dp*-expression D having a quantifier, I use the notation $S[(D)_n]$ to refer to a sentence S having a designated

Figure 2: The meta-rules for N_0^+

$$\frac{}{\Gamma, S \vdash S}\ (Ax)$$

$$\frac{\Gamma, [\mathbf{j}\ \mathsf{isa}\ X]_i \vdash S[\mathbf{j}]}{\Gamma \vdash S[(\mathsf{every}\ X)_{r(S[\mathbf{j}])+1}]}\ (eI^i)$$

$$\frac{\Gamma \vdash \mathbf{j}\ \mathsf{isa}\ X \quad \Gamma \vdash S[\mathbf{j}]}{\Gamma \vdash S[(\mathsf{some}\ X)_{r(S[\mathbf{j}])+1}]}\ (sI)$$

$$\frac{\Gamma \vdash S[(\mathsf{every}\ X)_{r(S[\mathbf{j}])+1}] \quad \Gamma \vdash \mathbf{j}\ \mathsf{isa}\ X \quad \Gamma, [S[\mathbf{j}]]_i \vdash S'}{\Gamma \vdash S'}\ (eE^i)$$

$$\frac{\Gamma \vdash S[(\mathsf{some}\ X)_{r(S[\mathbf{j}])+1}] \quad \Gamma, [\mathbf{j}\ \mathsf{isa}\ X]_j, [S[\mathbf{j}]]_i \vdash S'}{\Gamma \vdash S'}\ (sE^{i,j})$$

where $\mathbf{j}$ is fresh for $\Gamma, S[\mathsf{every}\ X]$ in (eI), and for $\Gamma, S[\mathsf{some}\ X], S'$ in (sE).

position filled by D, where n is the *scope level (sl)* of the quantifier in D. In case D has no quantifier (i.e., it is a parameter), *sl* $= 0$. The higher *sl*, the higher the scope. For example, $S[(\mathsf{every}\ X)_1]$ schematizes a sentence S with a designated occurrence of $\mathsf{every}\ X$ of the lowest scope. An example of a higher scope is $S[(\mathsf{some}\ X)_2]$, having $\mathsf{some}\ X$ in the higher scope, like in the object wide-scope reading of $(\mathsf{every}\ X)_1$ loves $(\mathsf{some}\ Y)_2$. I use the conventions that within a rule, both $S[D_1]$ and $S[D_2]$ refer to the *same* designated position in S, and when the *sl* can be unambiguously determined it is omitted. I use $r(S)$ to indicate the *rank* of S, the highest *sl* on a *dp* within S. Note that for a ground S, $r(S) = 0$.

Recall that in a rule, the notation $[\cdots]_i$ indicates an assumption *discharged* by an application of that rule. The indices of the assumptions discharged by a rule appear as superscripts on the rule name. The usual notion of (tree-shaped) derivation is assumed. I again use $\mathcal{D}$ to range over derivations, where $\mathcal{D}^{\Gamma \vdash S}$ is a derivation of sentence S from assumptions Γ. I use Γ, S for extending Γ with a sentence S. A more detailed explanation of the rules is presented in (Francez and Dyckhoff 2010). However, it is evident that all quantification in the fragment is restricted. In addition to this restriction I will add QDR in the next section.

The following is a convenient *derived* E-rule, that will be used to shorten derivations.

$$\frac{\Gamma \vdash S[(\mathsf{every}\ X)_{r(S[\mathbf{j}])+1}] \quad \Gamma \vdash \mathbf{j}\ \mathsf{isa}\ X}{\Gamma \vdash S[\mathbf{j}]}\ (e\hat{E})$$

Below is an example derivation establishing

$$\text{some } U \text{ isa } X,\ (\text{every } X)_2\ R\ (\text{some } Y)_1,\ \text{every } Y \text{ isa } Z\ \vdash\ (\text{some } U)_1\ R\ (\text{some } Z)_2.$$

Let $\overset{\mathscr{D}_1}{(\text{some } U)_2\ R\ (\text{some } Z)_1}$ and $\overset{\mathscr{D}_2}{(\text{some } U)_1\ R\ (\text{some } Y)_2}$ be the following two sub-derivations.

$$\mathscr{D}_1 :\quad \dfrac{\text{some } U \text{ isa } X \qquad \dfrac{[\mathbf{r} \text{ isa } U]_1 \qquad \dfrac{(\text{every } X)_2\ R\ (\text{some } Y)_1 \quad [\mathbf{r} \text{ isa } X]_2}{\mathbf{r}\ R \text{ some } Y}\,(e\hat{E})}{(\text{some } U)_2\ R\ (\text{some } Y)_1}\,(sI)}{(\text{some } U)_2\ R\ (\text{some } Y)_1}\,(sE^{1,2})$$

$$\mathscr{D}_2 :\quad \dfrac{[\text{some } U\ R\ \mathbf{j}]_3 \qquad \dfrac{\text{every } Y \text{ isa } Z \quad [\mathbf{j} \text{ isa } Y]_4}{\mathbf{j} \text{ isa } Z}\,(e\hat{E})}{(\text{some } U)_1\ R\ (\text{some } Z)_2}\,(sI)$$

The whole derivation combines the two sub-derivations by

$$\dfrac{\overset{\mathscr{D}_1}{(\text{some } U)_2\ R\ (\text{some } Y)_1} \qquad \overset{\mathscr{D}_2}{(\text{some } U)_1\ R\ (\text{some } Z)_2}}{(\text{some } U)_1\ R\ (\text{some } Z)_2}\,(sE^{3,4})$$

For a derivation $\mathscr{D}$ of S, its *root* is given by $\rho(\mathscr{D}) = S$. This function is extended to collections of derivations Δ by $\rho(\Delta) = \{\rho(\mathscr{D}) \mid \mathscr{D} \in \Delta\}$, and further extended to contextualized functions $\mathscr{F}$ by $\rho(\mathscr{F}) = \cup_{\Gamma}\rho(\mathscr{F}(\Gamma))$.

In order to understand better the PTS of E_0^+, consider one of its well-known features: *quantifier scope ambiguity*. The following E_0^+ sentences are usually attributed to two readings each, with the following FOL-expressions of their respective truth-conditions in model-theoretic semantics.

$$\text{Every girl loves some boy} \tag{3.34}$$

$$\text{Some girl loves every boy} \tag{3.35}$$

Consider sentence (3.34).

Subject wide-scope (sws): $\forall x.\mathbf{girl}(x) \rightarrow \exists y.\mathbf{boy}(y) \wedge \mathbf{love}(x, y)$

Subject narrow-scope (sns): $\exists y.\mathbf{boy}(y) \wedge \forall x.\mathbf{girl}(x) \rightarrow \mathbf{love}(x, y)$

In the proposed PTS, the difference in meanings reflects itself by the two readings having *different uses of the grounds for assertion*. This is manifested in derivations by different *orders of introduction* of the subject and object *dps*. Following Moss (2010), I disambiguate ambiguous sentences taking part in derivations.

Subject wide-scope (sws):

$$\dfrac{\dfrac{\dfrac{\dfrac{[\mathbf{r}\ \text{isa girl}]_i}{\mathscr{D}_1}}{\mathbf{r}\ \text{loves}\ \mathbf{j}} \qquad \dfrac{\mathscr{D}_2}{\mathbf{j}\ \text{isa boy}}}{\mathbf{r}\ \text{loves (some boy)}_1}\,(sI)}{\text{(every girl)}_2\ \text{loves (some boy)}_1}\,(eI^i) \tag{3.36}$$

Subject narrow-scope (sns):

$$\dfrac{\dfrac{\dfrac{\dfrac{[\mathbf{r}\ \text{isa girl}]_i}{\mathscr{D}_1}}{\mathbf{r}\ \text{loves}\ \mathbf{j}}}{\text{(every girl)}_1\ \text{loves}\ \mathbf{j}}\,(eI^i) \qquad \dfrac{\mathscr{D}_2}{\mathbf{j}\ \text{isa boy}}}{\text{(every girl)}_1\ \text{loves (some boy)}_2}\,(sI) \tag{3.37}$$

Note that there is no way to introduce a *dp* with a narrow-scope where the *dp* with the wider-scope has already been introduced. In the N_0^+ calculus, only disambiguated sentences participate.

3.1.2 Relative clauses and intersective adjectives

I next add relative clauses to the fragment, followed by intersective adjectives. This fragment transcends the locality of subcategorization in E_0^+, in having *long-distance dependencies*. It also has unbounded number of adjectival modifications. I refer to this (still positive) fragment as E_1^+. Note that, in contrast to E_0^+, E_1^+ is infinite. Typical sentences include the following.

every some/boy loves every/some girl
who(m) smiles/loves every/some flower/whom some girl loves (3.38)

every/some girl is a girl who loves every/some boy (3.39)

some boy loves every/some girl who loves every boy who smiles
(*nested relative clause*) (3.40)

So, girl who smiles and girl who loves every boy are *compound nouns*. I treat somewhat loosely the issue of the case of the relative pronoun, in the form of who(m), abbreviating either *who* or *whom*, as the case requires. I extend the notation with $S[-]$, which denotes, for S including a parameter in some distinguished position, the result of removing that parameter, leaving that position unoccupied. Examples are loves every girl (a parameter removed from subject position in **j** loves every girl), and every girl loves (a parameter removed from object position in every girl loves **k**).

The corresponding ND-system N_1^+ extends N_0^+ by adding the following I/E-rules.

$$\frac{\Gamma \vdash \mathbf{j} \text{ isa } X \quad \Gamma \vdash S[\mathbf{j}]}{\Gamma \vdash \mathbf{j} \text{ isa } X \text{ who } S[-]} \ (relI)$$
$$\frac{\Gamma \vdash \mathbf{j} \text{ isa } X \text{ who } S[-] \quad \Gamma, [\mathbf{j} \text{ isa } X]_i, [S[\mathbf{j}]]_j \vdash S'}{\Gamma \vdash S'} \ (relE^{i,j}) \tag{3.41}$$

The simplified derived E-rules are:

$$\frac{\Gamma \vdash \mathbf{j} \text{ isa } X \text{ who } S[-]}{\Gamma \vdash \mathbf{j} \text{ isa } X} \ (rel\hat{E})_1 \qquad \frac{\Gamma \vdash \mathbf{j} \text{ isa } X \text{ who } S[-]}{\Gamma \vdash S[\mathbf{j}]} \ (rel\hat{E})_2 \tag{3.42}$$

The familiar conjunctive behavior of relative clauses is exhibited here by its rules, resembling the rules for logical conjunction.

As an example of a derivation in this fragment, consider

$$\text{some girl who smiles sings} \vdash_{N_1^+} \text{some girl sings} \tag{3.43}$$

exhibiting the *upward monotonicity* of some in its first argument.

$$\frac{\text{some } X \text{ who } P_1 \ P_2 \qquad \dfrac{\dfrac{[\mathbf{r} \text{ isa } X \text{ who } P_1]_1}{\mathbf{r} \text{ isa } X} \ (rel\hat{E})_1 \quad [\mathbf{r}\ P_2]_2}{\text{some } X \ P_2} \ (sI)}{\text{some } X \ P_2} \ (sE^{1,2}) \tag{3.44}$$

Finally, I augment E_1^+ with sentences containing *adjectives*, schematized by A. I consider here only what is known in model-theoretic semantics as *intersective adjectives*. Typical sentences are:

Some girl is a beautiful girl/clever beautiful girl/clever beautiful red-headed girl (3.45)

every/some beautiful girl smiles (3.46)

$$\text{every/some beautiful girl loves every/some clever boy} \tag{3.47}$$

A noun preceded by an adjective is again a (compound) noun (the syntax is treated more precisely once the grammar is presented, as in Francez *et al.* 2010). Denote this extension still by E_1^+. Recall that in the N_1^+ rules, the noun schematization should be taken over compound nouns too. Note that I augment N_1^+ with the following ND-rules for adjectives.

$$\begin{gathered}\frac{\Gamma \vdash \mathbf{j}\ \mathsf{isa}\ X \quad \Gamma \vdash \mathbf{j}\ \mathsf{is}\ A}{\Gamma \vdash \mathbf{j}\ \mathsf{isa}\ A\, X}\ (adjI) \\ \frac{\Gamma \vdash \mathbf{j}\ \mathsf{isa}\ A\, X \quad \Gamma, [\mathbf{j}\ \mathsf{isa}\ X]_1, [\mathbf{j}\ \mathsf{is}\ A]_2 \vdash S'}{\Gamma \vdash S'}\ (adjE^{1,2})\end{gathered} \tag{3.48}$$

Again, the following *derived* E-rules are used to shorten presentations of example derivations.

$$\frac{\Gamma \vdash \mathbf{j}\ \mathsf{isa}\ A\, X}{\Gamma \vdash \mathbf{j}\ \mathsf{isa}\ X}\ (adj\hat{E}_1) \quad \frac{\Gamma \vdash \mathbf{j}\ \mathsf{isa}\ A\, X}{\Gamma \vdash \mathbf{j}\ \mathsf{is}\ A}\ (adj\hat{E}_2) \tag{3.49}$$

Note that the intersectivity here is manifested by the rules themselves (embodying an invisible conjunctive operator) at the sentential level. These rules induce intersectivity as a lexical property of (some) adjectives by the way lexical meanings are extracted from sentential meanings, as shown in (Francez *et al.* 2010).

The following sequent, the corresponding entailment of which is often taken as the definition of intersective adjectives, is derivable in N_1^+:

$$\mathbf{j}\ \mathsf{isa}\ A\, X,\ \mathbf{j}\ \mathsf{isa}\ Y \vdash \mathbf{j}\ \mathsf{isa}\ A\, Y \tag{3.50}$$

as shown by

$$\frac{\mathbf{j}\ \mathsf{isa}\ Y \quad \dfrac{\mathbf{j}\ \mathsf{isa}\ A\, X}{\mathbf{j}\ \mathsf{is}\ A}\ (adj\hat{E}_2)}{\mathbf{j}\ \mathsf{isa}\ A\, Y}\ (adjI) \tag{3.51}$$

As an example of derivations using the rules for adjectives, consider the following derivation for

$$\mathbf{j}\ \text{loves every girl} \ \vdash\ \mathbf{j}\ \text{loves every beautiful girl} \tag{3.52}$$

In model-theoretic semantics terminology, the corresponding entailment is a witness to the *downward monotonicity* of the meaning of every in its second argument. I use an obvious schematization.

$$\dfrac{\dfrac{\mathbf{j}\ R \text{ every } Y \qquad \dfrac{[\mathbf{r} \text{ isa } A\ Y]_1}{\mathbf{r} \text{ isa } Y}\,(adj\hat{E})}{\mathbf{j}\ R\ \mathbf{r}}\,(e\hat{E})}{\mathbf{j}\ R \text{ every } A\ Y}\,(eI^1) \tag{3.53}$$

Under this definition of the meaning of intersective adjectives, such adjectives are also *extensional*, in the sense of satisfying the following entailment:

$$\text{every } X \text{ isa } Y \ \vdash\ \text{every } A\,X \text{ isa } A\ Y \tag{3.54}$$

as shown by the following derivation:

$$\dfrac{\dfrac{\dfrac{\text{every } X \text{ isa } Y \qquad \dfrac{[\mathbf{j} \text{ isa } A\,X]_1}{\mathbf{j} \text{ isa } X}\,(adj\hat{E}_1)}{\mathbf{j} \text{ isa } Y}\,(e\hat{E}) \qquad \dfrac{[\mathbf{j} \text{ isa } A\,X]_1}{\mathbf{j} \text{ is } A}\,(adj\hat{E}_2)}{\mathbf{j} \text{ isa } A\ Y}\,(adjI)}{\text{every } A\,X \text{ isa } A\ Y}\,(eI^1) \tag{3.55}$$

The proof of harmony of N_1^+ can be found in (Francez and Dyckhoff 2010) and is not repeated here.

3.1.3 Sentential meanings

Again, a derivation is *canonical* if it essentially ends with an application of an I-rule; I use $\vdash^c$ for canonical derivability, denote by $[\![S]\!]^{\mathrm{c}}_{\Gamma}$ the collection of canonical derivations of S from Γ, and by $[\![S]\!]^{*}_{\Gamma}$ the collection of all derivations of S from Γ.

Those proof-theoretic collections are used to define meanings. Note that these are strictly proof-theoretic denotations, independent from any notion of a model, entities, and the like.

Definition 3.6 (PTS-meaning, semantic values)

1. For a non-ground $S \in L_1^+$, its *(reified) meaning* (also referred to as its *contributed* semantic value) is given by $[\![S]\!] \overset{\text{df.}}{=} \lambda\Gamma.[\![S]\!]^{\mathrm{c}}_{\Gamma}$ $[= \lambda\Gamma.\{\mathcal{D}^{\Gamma\vdash^{c}S}\}]$.

   Recall that for a *ground* S, $[\![S]\!]$ is assumed *given*. The meaning of non-ground pseudo-sentences (and E_0^+-sentences in particular) is defined *relative* to the given meanings of ground pseudo-sentences.

2. For an arbirary $S \in L_1^+$, its *contributing semantic value* is given by $[\![S]\!]^* \stackrel{\text{df.}}{=} \lambda\Gamma.[\![S]\!]_\Gamma^*$.

This distinction corresponds to the one that Dummett (1993, p. 48) introduced between *assertoric content* and *ingredient sense*. The content of an (affirmative) sentence S is the meaning of S in isolation, on its own. The ingredient sense of S is what S contributes to the meaning of any S' in which S occurs as a sub-expression, a component. This distinction is propagated to sub-sentential phrases as well. I will be concerned here with the contents of sentences only.

The main characteristic of this definition of (proof-theoretic) meaning is the notion of entailment it induces. A more comprehensive discussion can be found in (Francez 2014c).

By defining sentential meanings in this way, I do not allude to any logical form of the sentence differing from its surface form. In accordance with many views in philosophy of language, every derivation in the meaning of a sentence S can be viewed as providing $G[\![S]\!]$, *grounds for asserting* S. Definition (2.4) is adapted to the current fragment.

Definition 3.7 (grounds for assertion – NL) For $S \in E_1^+$, $G[\![S]\!] \stackrel{\text{df.}}{=} \{\Gamma \mid \Gamma \vdash^c S\}$, where Γ consists of E_1^+-sentences only. Parameters are not observable in grounds of assertion.

The refinement of the (reified) sentential meanings via '$\equiv_G$' is used here too, for the CIP (see below). A more comprehensive discussion of extensions of the fragment and some technicalities accompany the original presentation of the PTS in (Francez and Dyckhoff 2010).

3.2 *Quantifier domain restriction*

In this section, I develop the proof-theoretic semantics for QDR in setting of the natural language fragment E_1^+. This setting is more suitable for that task than that of FOL and FOL$_{rq}$, that were considered for ease of presentation of the approach, being more familiar to most readers than the dedicated N_1^+.

Definition 3.8 (NLDR-context) An *NLDR-context* (NL domain restricting context) c is a finite collection Γ_c of pseudo-sentences with one parameter only, where $\Gamma_{c,\mathbf{j}}$ is the sub-collection with the parameter $\mathbf{j}$.[13] Let *CNL* (NL contexts) be the collection of all NLDR-contexts.

[13] For simplicity, I assume this sub-collection is a singleton.

The NLDR-contexts $\Gamma_{c,\mathbf{j}}(\mathbf{j})$ can be of one of the following forms: $\mathbf{j}$ isa X (X is a noun), $\mathbf{j}$ is A (where A is an adjective) or $\mathbf{j}$ P (where P is a verb phrase). Note that compound contextual restrictions can also be imposed, as, for example, in $\Gamma_{c,\mathbf{j}} = \mathbf{j}$ $\mathsf{isa\ man\ whom\ every\ girl\ loves}$. Since the fragment E_1^+ has only modification by means of (intersective) adjectives and relative clauses, all the examples will be restricted to such modification. Extensions, for example, to incorporate preposition phrases, are not an obstacle in principle, but none have been proposed yet.

Restricting universal quantification: Again, an NLDR-context c provides a *discharged* assumption for imposing its restriction.

$$\frac{\Gamma, [\mathbf{j}\ \mathsf{isa}\ X]_i, [\Gamma_{c,\mathbf{j}}(\mathbf{j})]_j \vdash_c S[\mathbf{j}]}{\Gamma \vdash_c S[(\mathsf{every}\ X)_{r(S[\mathbf{j}])+1}]}\ (eI^{i,j}_{CNL}),\ \mathbf{j}\text{ fresh for }\Gamma \tag{3.56}$$

$$\frac{\Gamma \vdash_c S[(\mathsf{every}\ X)_{r(S[\mathbf{j}])+1}] \quad \Gamma \vdash_c \mathbf{k}\ \mathsf{isa}\ X \quad \Gamma \vdash_c \Gamma_{c,\mathbf{j}}(\mathbf{k}) \quad \Gamma, [S[\mathbf{k}]]_i \vdash_c S'}{\Gamma \vdash_c S'}\ (eE^{i}_{CNL})$$

$$\mathbf{k}\text{ fresh} \tag{3.57}$$

Again, a family of I/E-rules is employed, for all possible NLDR-contexts.

Example 3.4 Below is a derivation establishing

$$\mathsf{every\ Italian\ woman\ smiles, every\ woman\ who\ smiles\ is\ beautiful} \\ \vdash_c\ \mathsf{every\ woman\ is\ beautiful} \tag{3.58}$$

in an NLDR-context c with $\Gamma_{c,\mathbf{k}}(\mathbf{k}) = \mathbf{k}$ $\mathsf{is\ Italian}$, intended to restrict the universal quantification on women to a universal quantification on Italian women. The observant reader will notice that Example 2.2 is a regimentation of this example in FOL. Since there is no quantifier scope ambiguity involved in this example, I omit in the derivation the scope level indicator to avoid notational clutter. Also, for typographical reasons, I abbreviate in the derivation Woman, Italian, Beautiful and Smiles to W, I, B and S, respectively.

$$\dfrac{\dfrac{\dfrac{[\mathbf{k}\ \textsf{isa W}]_1 \quad \dfrac{\dfrac{[\mathbf{k}\ \textsf{is I}]_2 \quad [\mathbf{k}\ \textsf{isa W}]_1}{\mathbf{k}\ \textsf{isa I W}}\,(adjI) \quad \textsf{every I W S}}{\mathbf{k}\ \textsf{S}}\,(e\hat{E})}{\mathbf{k}\ \textsf{isa W who S}}\,(relI) \quad \textsf{every W who S is B}}{\mathbf{k}\ \textsf{is B}}\,(e\hat{E})}{\textsf{every W is B}}\,(eI^{1,2}_{CNL}) \tag{3.59}$$

It is interesting to note that QDR holds, no matter which scope level the restricted quantifier is in.

Example 3.5 Consider the following scope variants of

$$\textsf{some man admires every actress} \tag{3.60}$$

In the NLDR-context c with $\Gamma_{c,\mathbf{k}} = \mathbf{k}$ is Italian, intended to restrict the universal quantification on actresses to universal quantification on Italian actresses. I use the abbreviations P for philosopher, M for man, I for Italian, A for actress, adm for admires and S for smart, to show that

$$\begin{aligned} \textsf{some } P \textsf{ isa } M, \textsf{every } P \textsf{ is } S, \textsf{every } S\ M \textsf{ adm every } I\ A \\ \vdash_c \textsf{some } M \textsf{ adm every } A \end{aligned} \tag{3.61}$$

under both scope variants of the conclusion. For typographical reasons, the derivations are presented with a common sub-derivation $\mathscr{D}$ factored out.

$$\mathscr{D} = \dfrac{\dfrac{[\mathbf{j}\ \textsf{isa } M]_1 \quad \dfrac{[\mathbf{j}\ \textsf{isa } P]_2 \quad \textsf{every } P \textsf{ is } S}{\mathbf{j}\ \textsf{is } S}\,(e\hat{E})}{\mathbf{j}\ \textsf{isa } S\ M}\,(adjI) \quad \textsf{every } S\ M \textsf{ adm every } I\ A}{\mathbf{j}\ \textsf{adm every } I\ A}\,(e\hat{E}) \tag{3.62}$$

Subject wide scope: The derivation is (with obvious abbreviations and Γ omitted):

$$\dfrac{\textsf{some } P \textsf{ isa } M \quad \dfrac{[\mathbf{j}\ \textsf{isa } M]_1 \quad \dfrac{\dfrac{\dfrac{[\mathbf{k}\ \textsf{is } I]_3 \quad [\mathbf{k}\ \textsf{isa } A]_4}{\mathbf{k}\ \textsf{isa } I\ A}\,(adjI) \quad \overset{\mathscr{D}}{\mathbf{j}\ \textsf{adm every } I\ A}}{\mathbf{j}\ \textsf{adm}\ \mathbf{k}}\,(e\hat{E})}{\mathbf{j}\ \textsf{adm (every } A)_1}\,(eI^{3,4}_{CNL})}{(\textsf{some } M)_2 \textsf{ adm (every } A)_1}\,(sI)}{(\textsf{some } M)_2 \textsf{ adm (every } A)_1}\,(sE^{1,2}) \tag{3.63}$$

Object wide scope: The derivation is

$$\dfrac{\textsf{some } P \textsf{ isa } M \quad \dfrac{[\mathbf{j} \textsf{ isa } M]_1 \quad \dfrac{\dfrac{[\mathbf{k} \textsf{ is } I]_3 \quad [\mathbf{k} \textsf{ isa } A]_4}{\mathbf{k} \textsf{ isa } I\ A}(adjI) \quad \dfrac{\mathscr{D}}{\mathbf{j} \textsf{ adm every } I\ A}}{\mathbf{j} \textsf{ adm } \mathbf{k}}(e\hat{E})}{(\textsf{some } M)_1 \textsf{ adm } \mathbf{k}}(sI)}{\dfrac{(\textsf{some } M)_1 \textsf{ adm } \mathbf{k}}{(\textsf{some } M)_1 \textsf{ adm } (\textsf{every } A)_2}(eI^{3,4}_{CNL})}(sE^{1,2}) \tag{3.64}$$

Example 3.6 The following example from (Stanley and Szabó 2000) is pointed out as being difficult for MTS-handling, as it seemingly requires context-shift during meaning evaluation.

$$\textsf{every sailor waved to every sailor} \tag{3.65}$$

where the context imposes the restriction that the quantification in the subject is restricted to one kind of sailors, say sailors on the ship, while the object quantification is restricted, say, to sailors on the shore. Under the current approach, such examples pose no problem whatsoever. Suppose that **j** is the parameter used to introduce every sailor in the subject, while **k** is the parameter used to introduce every sailor in the object (where both scope relations are equivalent). Then, all we have to do is consider a context $c_{sailors}$, with $\Gamma_{c_{sailors},j} = \mathbf{j}\ \textsf{is}-\textsf{on}-\textsf{the}-\textsf{ship}$, and $\Gamma_{c_{sailors},k} = \mathbf{k}\ \textsf{is}-\textsf{on}-\textsf{the}-\textsf{shore}$. No context shift is involved. As a full derivation is somewhat lengthy, I skip the details.

Restricting existential quantification:

$$\frac{\Gamma \vdash \mathbf{j} \textsf{ isa } X \quad \Gamma \vdash \Gamma_{c,\mathbf{j}}(\mathbf{j}) \quad \Gamma \vdash S[\mathbf{j}]}{\Gamma \vdash_c S[(\textsf{some } X)_{r(S[\mathbf{j}])+1}]}(sI_{CNL}) \tag{3.66}$$

$$\frac{\Gamma \vdash S[(\textsf{some } X)_{r(S[\mathbf{j}])+1}] \quad \Gamma, [\mathbf{k} \textsf{ isa } X]_i, [\Gamma_{c,\mathbf{j}}(\mathbf{k})]_j, [S[\mathbf{k}]]_k \vdash S'}{\Gamma \vdash S'}(sE^{i,j,k}_{CNL}) \tag{3.67}$$

where **k** is fresh for $\Gamma, S[\textsf{some } X], S'$ in (sE_{CNL}).

The reductions needed to show the harmony of the *CNL*-rules are very similar to those for the regular rules (shown in Francez and Dyckhoff 2010) and are omitted.

Next, I show how the CIP is satisfied for E_1^+. Note that the fragment includes neither implication nor conjunction (on the sentential level).

To express the CIP effect, I use the following notation. For $S[(q\ X)]$ (with q either every or some), let $S_{\Gamma_{c,\mathbf{j}}}$ be defined as

$$S_{\Gamma_{c,\mathbf{j}}} = \begin{cases} S[(q\ X\ \textsf{who isa}\ Y)] & \Gamma_{c,\mathbf{j}}(\mathbf{j}) = \mathbf{j}\ \textsf{isa}\ Y \\ S[(q\ A\ X)] & \Gamma_{c,\mathbf{j}}(\mathbf{j}) = \mathbf{j}\ \textsf{is}\ A \\ S[(q\ X\ \textsf{who}\ P)] & \Gamma_{c,\mathbf{j}}(\mathbf{j}) = \mathbf{j}\ P \end{cases} \tag{3.68}$$

Theorem 3.2 (E_1^+ **context incorporation**)

$$\Gamma \vdash_c S[(q\ X)] \text{ iff } \Gamma \vdash S_{\Gamma_{c,\mathbf{j}}} \tag{3.69}$$

Proof: I will show only the proof of the first case, for $q = \textsf{every}$; all other cases are similar. To simplify, I also omit the scope indications.

1. Asume $\Gamma \vdash_c S[(\textsf{every}\ X)]$, where $\Gamma_{c,\mathbf{j}} = \mathbf{j}\ \textsf{isa}\ Y$. So, the derivation ends with (omitting scope indication)

$$\dfrac{\Gamma, [\mathbf{j}\ \textsf{isa}\ X]_i, [\mathbf{j}\ \textsf{isa}\ Y]_j \vdash_c S[\mathbf{j}]}{\Gamma \vdash_c S[(\textsf{every}\ X)]}\,(eI^{i,j}_{CNL}),\ \mathbf{j} \text{ fresh for } \Gamma \tag{3.70}$$

   Therefore, the following derivation can be formed, where the induction hypothesis on the premise uses '$\vdash$' instead of '$\vdash_c$'.

$$\dfrac{\dfrac{\dfrac{\Gamma, [\mathbf{j}\ \textsf{isa}\ X\ \textsf{who isa}\ Y]_i}{\Gamma \vdash \mathbf{j}\ \textsf{isa}\ X}\,(relE) \quad \dfrac{\Gamma, [\mathbf{j}\ \textsf{isa}\ X\ \textsf{who isa}\ Y]_i}{\Gamma \vdash \mathbf{j}\ \textsf{isa}\ Y}\,(relE)}{\Gamma \vdash S[\mathbf{j}]}\,(ass.)}{S[(\textsf{every}\ X\ \textsf{who isa}\ Y)]}\,(eI^i) \tag{3.71}$$

2. Assume $\Gamma \vdash S[(\textsf{every}\ X\ \textsf{who isa Y})]$. The derivation (again, omitting scope indication) ends with

$$\dfrac{\Gamma, [\mathbf{j}\ \textsf{isa}\ X\ \textsf{who isa}\ Y]_i \vdash S[\mathbf{j}]}{\Gamma \vdash S[(\textsf{every}\ X\ \textsf{who isa Y})]}\,(eI^i) \tag{3.72}$$

   Let $\Gamma_{c,\mathbf{j}} = \mathbf{j}\ \textsf{isa}\ Y$. Therefore, the following derivation can be formed:

$$\dfrac{\dfrac{\dfrac{[\Gamma \vdash \mathbf{j}\ \textsf{isa}X]_i \quad [\Gamma \vdash \mathbf{j}\ \textsf{isa}Y]_j}{\Gamma \vdash \mathbf{j}\ \textsf{isa}\ X\ \textsf{who isa}\ Y}\,(relI)}{S[\mathbf{j}]}\,(ass.)}{\Gamma \vdash_c S[(\textsf{every}\ X)]}\,(eI^{i,j}_{CNL}) \tag{3.73}$$

Once again, by inspecting the rules, we obtain that

$$[\![S[(q\ X)]]\!]_c = [\![S_{\Gamma_{c,j}}]\!] \tag{3.74}$$

validating the CIP. Note again the correspondence between derivations, where a use of (qI_{NLC}) is associated with (qI) followed by (*rel I*) (or by (*Adj I*) in some of the cases), and similarly for the E-rules.

4 CONCLUSIONS

The paper introduces proof-theoretic semantics for contextual domain restriction as an alternative to the model-theoretic meaning generally found in the literature. In addition to providing yet another example for the feasibility of PTS for natural language meanings, the paper points to an advantage (in my opinion) of the PTS approach to QDR over the MTS approach; namely, the CIP principle, by which every contextually restricted quantified sentence has the same meaning as a context-independent variant thereof, where the contextual restriction is incorporated as a phrase in the sentence. Thus, no equivalent of intersection with *arbitrary* subsets of the quantification domain, not being the denotation (in the model) of any NL expression, is involved. Some other advantages related to multiple quantification have also been shown. In particular, *both* facets of the QDR-problem pointed out by Stanley and Szabó (2000), namely the descriptive and the fundamental, are treated, in contrast to the MTS discussion in the literature, which usually evades the latter.

As observed by one of the referees of this paper, an important phenomenon related to contextual meaning variation, namely, pronominal *binding*, is not covered by the proposed PTS. The reason is that currently the fragment for which a PTS has been proposed does not include pronouns at all. I consider this to be a topic of further work, both extending the fragment with pronouns and investigating the impact of such an extension of the general contextual QDR-problem.

The approach was also exemplified in two variants of FOL (first-order logic). The current interface between the contextual restriction and the sentential derivation is through the name of the variable involved (or in the NL case, through the parameter). This might seem somewhat ad hoc, and a more transparent binding of contextual restrictions and the corresponding quantifiers should be sought.

REFERENCES

Hanoch BEN-YAMI (2006), A critique of Frege on common nouns, *Ratio*, 19(2):148–155.

Robert B. BRANDOM (2000), *Articulating reasons*, Harvard University Press, Cambridge, MA.

Michael DUMMETT (1993), *The logical basis of metaphysics*, Harvard University Press, Cambridge, MA, USA, hard copy 1991.

Nissim FRANCEZ (2014a), The granularity of meaning in proof-theoretic semantics, in Nicholas ASHER and Sergei SOLOVIEW, editors, *Proceedings of the 8th International Conference on Logical Aspects of Computational Linguistics (LACL), Toulouse, France, June 2014*, volume 8535 of *LNCS*, pp. 96–106, Springer Verlag, Berlin/Heidelberg, Germany.

Nissim FRANCEZ (2014b), A logic inspired by natural language: quantifiers as subnectors, *Journal of Philosophical Logic*, doi:10.1007/s10992-014-9312-z.

Nissim FRANCEZ (2014c), Views of proof-theoretic semantics: Reified proof-theoretic meanings, *Journal of Computational Logic*, special issue in honour of Roy Dyckhoff, doi:10.1093/logcom/exu035.

Nissim FRANCEZ and Gilad BEN-AVI (2011), Proof-theoretic semantic values for logical operators, *Review of Symbolic Logic*, 4(3):337–485.

Nissim FRANCEZ and Gilad BEN-AVI (2014), A proof-theoretic reconstruction of generalized quantifiers, *Journal of Semantics*, doi:10.1093/jos/ffu001.

Nissim FRANCEZ and Roy DYCKHOFF (2010), Proof-theoretic semantics for a natural language fragment, *Linguistics and Philosophy*, 33(6):447–477.

Nissim FRANCEZ and Roy DYCKHOFF (2012), A note on harmony, *Journal of Philosophical Logic*, 41(3):613–628.

Nissim FRANCEZ, Roy DYCKHOFF, and Gilad BEN-AVI (2010), Proof-theoretic semantics for subsentential phrases, *Studia Logica 94*, pp. 381–401.

Nissim FRANCEZ and Bartosz WIECKOWSKI (2014), A proof-theoretic semantics for contextual definiteness, in Enrico MORICONI and Laura TESCONI, editors, *Second Pisa Colloquium in Logic, Language and Epistemology*, ETS, Pisa, Italy.

Gottlob FREGE (1884), *Die Grundlagen der Arithmetik [The basic laws of arithmetics]*, Georg Olms, Hildesheim, Germany.

Gerhard GENTZEN (1969), Investigations into logical deduction, in M.E. SZABO, editor, *The collected papers of Gerhard Gentzen*, pp. 68–131, North-Holland, Amsterdam, Netherlands, English translation of the 1935 paper in German.

Michael GLANZBERG (2006), Context and unrestricted quantification, in Augustìne RAYO and Gabriel UZQUIANO, editors, *Absolute Generality*, Clarendon Press, Oxford, UK.

Michael MOORTGAT (1997), Categorial type logics, in Johan VAN BENTHEM and Alice TER MEULEN, editors, *Handbook of Logic and Language*, pp. 93–178, North-Holland, Amsterdam, Netherlands.

Lawrence MOSS (2010), Syllogistic logics with verbs, *Journal of Logic and Computation*, 20(4):947–967.

Francis Jeffrey PELLETIER (2003), Context dependence and compositionality, *Mind & Language*, 18(2):148–161.

Dag PRAWITZ (1965), *Natural deduction: A proof-theoretical study*, Almqvist and Wicksell, Stockholm, Sweden, soft cover edition by Dover, 2006.

Dag PRAWITZ (2006), Meaning approached via proofs, *Synthese*, 148:507–524.

Arthur N. PRIOR (1960), The runabout inference-ticket, *Analysis*, 21:38–39.

Peter SCHROEDER-HEISTER (1984), A natural extension of natural deduction, *Journal of Symbolic Logic*, 49:1284–1300.

Jason STANLEY and Zoltán Gendler SZABÓ (2000), On quantifier domain restriction, *Mind & Language*, 2-3:219–261.

Neil TENNANT (1997), *The taming of the true*, Oxford University Press, Oxford, UK.

Jan VON PLATO (2000), A problem with normal form in natural deduction, *Mathematical Logic Quarterly*, 46:121–124.

Jan VON PLATO (2001), Natural deduction with general elimination rules, *Archive for Mathematical Logic*, 40:541–567.

Dag WESTERSTÅHL (1985), Determiners and context sets, in Johan VAN BENTHEM and Alice TER MEULEN, editors, *Generalized Quantifiers in Natural Language*, Foris, Dordrecht, Netherlands.

Handling equivalence classes of Optimality-Theoretic comparative tableaux

Igor Yanovich
Universität Tübingen

ABSTRACT

Many Optimality-Theoretic tableaux contain exactly the same information, and equivalence-preserving operations on them have been an object of study for some two decades. This paper shows that several of the operations proposed in the earlier literature together are actually enough to express *any* possible equivalence-preserving transformation. Moreover, every equivalence class of comparative tableaux (equivalently, of sets of Elementary Ranking Conditions, or ERC sets) has a unique and computable normal form that can be derived using those elementary operations in polynomial time. Any equivalence-preserving operation on comparative tableaux (ERC sets) is thus computable, and normal form tableaux may therefore represent their equivalence classes without loss of generality.

Keywords: Optimality Theory, ERC set, tableau equivalence, normal form, equivalence-preserving transformations

Optimality Theory (OT) is a grammatical formalism based on constraint competition, formulated by Prince and Smolensky (1993) (later published as Prince and Smolensky (2004)). OT is especially popular in phonology, and is used to some extent in other branches of linguistics. In OT, a set of competing output forms $\{Output_1, Output_2, \ldots\}$ is generated by machine **Gen** for the underlying form *Input*. Each pair $\langle Input, Output_N \rangle$ is then evaluated against a set of constraints **Con**. The grammar of a particular language is modeled as an ordering of the universal set of constraints **Con** which determines the winning input-output pair for each *Input*: an input-output pair $\alpha = \langle Input, Output_N \rangle$ wins over another pair $\beta = \langle Input, Output_M \rangle$ when α incurs fewer violations than β in the most highly ranked constraint where α and β differ. The input-output pairs that do not lose to any other pair are declared grammatical.

The OT formalism expresses two important intuitions regarding how languages might function. First, it easily captures conditions of the form "try A; if impossible, try B; if also impossible, resort to C", which seem to frequently occur in natural language. Second, OT allows for elegant modeling of cross-linguistic variation and language change in terms of re-ranking of a universal set of constraints.

The information that a given dataset contributes constrains the possible rankings of constraints. Such information may be represented in the form of a comparative tableau (Prince 2000) or the corresponding set of Elementary Ranking Conditions, or ERC set (Prince 2002) . In this paper, I present an incremental step completing the development of a full theory of equivalence classes of comparative OT tableaux, or, equivalently, ERC sets.

Earlier work, especially that of Hayes (1997), Prince (2000), Prince (2002), Brasoveanu and Prince (2011)[1], and Prince (2006), has established a number of results concerning how one may transform the information in an OT tableau without loss. What has not yet been done in this line of research is to establish the limits of operations that preserve equivalence. For example, the following natural question has not been answered: given two arbitrary comparative tableaux or ERC sets, can we determine whether they contain identical information?[2]

The present paper fills this gap: I show that any (finite) comparative tableau may be (computably, and actually quite efficiently) transformed into a normal form, which is unique for the whole equivalence class. Moreover, this transformation is possible by applying a sequence of a set of five elementary operations and their inverses

[1] An earlier version (Brasoveanu and Prince 2005) was circulated through Rutgers Optimality Archive (ROA) `http://roa.rutgers.edu/`

[2] For a finite constraint set, there is only a finite number of possible rankings, so strictly speaking, brute-force testing for equivalence is possible: one may simply build every possible ranking and test whether the two tableaux/ERC sets are compatible with it. However, the number of logically possible rankings of n constraints is $n!$, so the complexity of brute-force testing is factorial in the number of constraints. This should be compared with the merely polynomial time complexity of our new test for equivalence through normalization given in Theorem (16).

already introduced in the literature. Only two of those are non-trivial, so a very small and simple set turns out to be sufficient to capture all the diversity of possible equivalence-preserving operations on tableaux. Normalization gives us a handle on equivalence classes of tableaux/ERC sets, as we show that each equivalence class contains exactly one normal form tableau. The normal form may therefore serve as the class's representative. A test for equivalence of arbitrary tableaux (computable for finite tableaux) involves normalizing the input tableaux and comparing the resulting normal form tableaux. The original tableaux are equivalent if and only if their normal forms are identical. Thanks to the normal form theorem proved in the present paper, the space of all possible equivalence-preserving operations may be enumerated, and the same is true of the members of which equivalence class.

1 INTRODUCTION

As a concrete example of how OT works, consider the pattern of final obstruent devoicing in Dutch.[3] Underlyingly, Dutch morphemes may have both voiced and voiceless obstruents: the morpheme for 'bed' is /bɛd/, surfacing faithfully in [bɛd-ən] 'beds', while the morpheme for 'dab' is /bɛt/, surfacing faithfully in [bɛt-ən] '(we) dab'. But when the final obstruent of either morpheme closes the syllable, it is realized on the surface by the same voiceless [t]: both 'bed' /bɛd/ and '(I) dab' /bɛt/ surface as [bɛt]. The following OT tableau demonstrates the violation patterns for several potential outputs corresponding to the underlying form /bɛd/:

(1)

UR: /bɛd/	*Voiced- Obs-Coda	Ident-Voice	*Voiced-Obs
a. [bɛd]	*		**
b. [bɛt]		*	*
c. [pɛd]	*	*	*
d. [pɛt]		**	

According to the OT conventions, solid vertical lines in the tableau indicate that the left-to-right order of the constraints corresponds to their ranking in the grammar: *Voiced-Obs-Coda ≫ Ident-Voice

[3] My description of the Dutch pattern is based on Kager (1999).

≫ *Voiced-Obs. The constraint *Voiced-Obs penalizes any voiced obstruent. Its specialized cousin *Voiced-Obs-Coda only penalizes voiced obstruents in the coda position of a syllable. Finally, Ident-Voice penalizes mismatches in voice between underlying and output consonants. The ranking in the tableau ensures that [bɛt] is the winning output form: [bɛd] and [pɛd] lose to [bɛt] in the highest constraint *Voiced-Obs-Coda, and [pɛt] loses to it in the next constraint Ident-Voice. Overall, the ranking says: “avoid voiced obstruents in the coda, but preserve them elsewhere”.[4] We worked through this example already knowing the ranking. Normally the work of an OT analyst proceeds in the opposite direction: she would know the constraints, the violation profiles, and the designated winner, and would need to uncover the ranking that selects the winner correctly. For that procedure, it is more convenient to use a comparative OT tableau, Prince (2000). The comparative counterpart of Tableau (1) is given in Tableau (2). Each row of a comparative tableau corresponds to a pair of the winner output and one of the loser outputs of the regular OT tableau as in Tableau (1). For a specific row corresponding to a specific winner-loser pair, if the winner incurs less violations than the loser in a given constraint, the relevant cell is marked with a W; if the loser incurs less violations, the cell is marked with an L. If there is a tie, it is marked with an *e*.

(2)

UR: /bɛd/	*Voiced-Obs-Coda	Ident-Voice	*Voiced-Obs
[bɛt]∼[bɛd]	W	L	W
[bɛt]∼[pɛd]	W	*e*	*e*
[bɛt]∼[pɛt]	*e*	W	L

It is easy to see that converting a traditional OT tableau into a comparative tableau loses information about the number of violations. But the lost information is irrelevant for recovering the ranking. Moreover, the characterization of rankings which select the correct winner becomes very simple with comparative tableaux: a ranking selects the right winner iff in every row, all L-constraints are dominated by a W-

[4] It is easy to check that the ranking in 1 predicts correct results for Dutch [bɛd-ən] ‘beds’, [bɛt-ən] ‘(we) dab’, and /bɛt/-[bɛt] ‘(I) dab’. It is also the only ranking selecting the correct winner in 1, though there exist tableaux whose winner can be correctly selected by more than one ranking.

constraint. A specific condition selecting the rankings compatible with a fixed row is called the Elementary Ranking Condition, or ERC, by Prince (2002). In Tableau (2), we can see for instance that the pair [bɛt]~[pɛt] necessitates the inclusion of a pairwise ranking IDENT-VOICE ≫ *VOICED-OBS into our grammar. On the other hand, another pair [bɛt]~[pɛd] does not add any useful information: without any Ls in the row, [pɛd] is going to lose to [bɛt] on any possible ranking of our three constraints (i.e., the ERC corresponding to this comparative row is trivial, as it is compatible with any ranking.) In what follows, I will be largely talking in terms of comparative rows and tableaux, but it is easy to translate this into talk about ERCs and ERC sets.

Turning to definitions, a **comparative tableau** is a possibly empty 2-dimensional matrix with labelled columns where each cell contains a W, an L or an *e*. The column labels of a given tableau form the **constraint set**. A **comparative row** is a comparative tableau with one row. The tableau with zero rows is special: it is compatible with any ranking whatsoever; we refer to it as $T_\top$. A **(total) ranking** is a total order of a constraint set. In what follows, we always assume that tableaux and rankings use the same fixed constraint set.

The following terminology, mostly borrowed from Prince (2002), will also be useful. A ranking M[5] is **(OT-)compatible** with a comparative tableau T iff for every row, every L-constraint is dominated by some W-constraint. We say that ranking M **covers** an L in constraint C in row r when M orders one of the W-constraints of r higher than the L-constraint C. We also say that a W in any constraint C' that dominates C under ranking M **covers** the L in C. If every ranking compatible with tableau T is also compatible with tableau U, we say that T **entails** U. When T and U are compatible with exactly the same rankings, they are called **OT-equivalent**. It is trivial to extend the notions to ERC sets.

Once a comparative tableau is computed, the actual input-output pairs are no longer needed for the task of determining the correct ranking. Thus we may freely combine several tableaux stemming from different input forms into a single bigger tableau: the input informa-

[5] Prince (2002) introduces the logical perspective on OT compatibility wherein rows/ERCs are formulas, and rankings are essentially models. Hence M, N as designations for rankings.

tion in it may be viewed as being about the grammar of the language rather than about particular linguistic forms. In this paper, we will be working exclusively with comparative tableaux.

Tableaux directly computed from particular linguistic forms are often suboptimal in how they represent information. For example, the second row of 2 may be omitted without any loss of information; similarly, the W in *VOICED-OBS in the first row is “false”, because replacing it with an *e* will not change which ranking selects the right winner. It thus becomes important to study equivalence relations between comparative tableaux/ERC sets. To name just a few examples, Hayes (1997) (cf. also a follow-up in Prince (2006)) seeks to find transformations for tableaux allowing for better information extraction; Prince (2000) introduces the notion of entailment between rows and tableaux; Brasoveanu and Prince (2011) define an algorithm transforming an arbitrary tableau into a small-size “basis” conveniently representing the same information.

The current paper continues that line of investigation. Namely, I prove that the equivalence-preserving operations introduced in the earlier literature are already enough to handle equivalence classes of comparative tableaux/ERC sets, once we add the necessary proofs. By definition, any (comparative) tableau T belongs to an equivalence class $\mathscr{C}$ such that any tableau in $\mathscr{C}$ is compatible with exactly the same rankings. Whenever there are such non-trivial equivalence classes, there is a problem of handling them: in geometry, there are congruence classes of geometrical figures; in proof theory, there often exist many proofs of the same statement; in lambda-calculus, there are plenty of equivalent lambda-terms. In all those cases we want to be able to obtain results common for the equivalence class. Our strategy for getting a handle on equivalent classes of OT tableaux will be fairly standard: we will find a special representative which exists in every equivalence class, and is unique in it — in other words, a normal form that can represent the class.

The plan is as follows. In Section 2, I review several elementary equivalence-preserving transformations of tableaux from the earlier literature, adding their inverses where needed. Later it will be shown that the introduced set of operations is functionally complete (that is, any equivalence-preserving transformation can be decomposed into a sequence of elementary transformations from the set). In Section 3 I define a normal form for OT tableaux, and prove the central result of the paper: a normal form is unique in its equivalence class. This means that the normal form may be used as the representative of a class, or its *name*. Finally, in Section 4 I provide several easy corollaries following from the normal form theorems. For example, we obtain a test of equivalence for OT tableaux, and a proof that bases of Brasoveanu and Prince (2011) are unique in their equivalence classes and thus can serve as class representatives (just as normal-form tableaux can).

2 FIVE ELEMENTARY EQUIVALENCE-PRESERVING TRANSFORMATIONS

In this section, we provide the definitions for five operations with inverses that will be shown in the next section to form a functionally complete set. The operations are either trivial (Operations (3) and (4)) or have been described and proven correct before (Operations (5) and (6) are either explicitly discussed by, or immediately follow from Prince (2002); Operation (7) is studied in Prince (2006)). The proofs of equivalence-preservation are provided here mainly for completeness' sake, so the readers familiar with the operations may wish to skip them. The novelty of the present paper is not in the operations themselves, but in the fact that together they form a functionally complete set that is enough to represent any possible equivalence-preserving operation whatsoever.

The order of columns in example tableaux below does *not* correspond to any ranking, unlike in the previous section.[6] Constraint names are chosen to be $C1$, $C2$, ..., rather than the usual meaningful names, to underscore the fact that the transformations are completely blind to actual linguistic content, and only concern the formal information encoded in a tableau.

We use variables M, N, ...for OT rankings; variables T, U, ...for comparative OT tableaux; and r and q for comparative OT rows. $W(r)$, for row r, denotes the set of constraints that have a W in r. Similarly for $L(r)$. This short notation allows us to define new rows compactly: e.g.,

[6] Sometimes the absence of order is marked by using dashed vertical lines. We refrain from this practice at the request of a reviewer.

if we say that $W(r) = \{C3\}$ and $L(r) = \{C1\}$, and Con is the 5-constraint set $\{C1, C2, C3, C4, C5\}$, then row r is the row (L, e, W, e, e).

The first two operations we will consider are trivial. First, row swaps defined in Operation (3) never affect OT-equivalence, as the order of the rows is not significant for determining whether a ranking M is compatible with the tableau. (If we think in terms of corresponding ERC sets, the very concept of row order becomes irrelevant.) Row swap is its all inverse. Second, if a tableau is not compatible with any ranking whatsoever (that is, if it puts contradictory requirements on the ranking of constraints), there is no useful information in it anyway, so as long as the tableau remains contradictory, any changes to it do not offend equivalence (Operation (4)).

(3) **Row swaps**: swapping any two rows preserves OT equivalence.

$C1$	$C2$	$C3$	$C4$
e	W	L	L
W	*e*	L	*e*

⇔

$C1$	$C2$	$C3$	$C4$
W	*e*	L	*e*
e	W	L	L

Proof: trivial.

(4) **Contradictory jumps**: for a contradictory tableau (that is, a tableau not compatible with any ranking), any row can be added, or, inversely, subtracted as long as the resulting tableau is still contradictory.

$C1$	$C2$	$C3$	$C4$
W	L	*e*	*e*
L	W	*e*	*e*

⇔

$C1$	$C2$	$C3$	$C4$
W	L	*e*	*e*
L	W	*e*	*e*
e	*e*	W	L

Proof: trivial.

Row splitting and its inverse, row merging, are also nearly trivial. Given the ERC theory of Prince (2002), it is easy to show that a row with several Ls is equivalent to a set of single-L rows. In ERC terms, such single-L rows have been called Primitive Ranking Conditions by Prince (2006, p. 4). The correctness of row splitting and row merging shows that covering each L in a multiple-L row is independent from covering the other Ls. Working with single-L rows, or PRCs, is often more convenient, especially when we turn all rows in a tableau into this single-L/PRC form.

(5) **Row splittings and mergings**: a row r is equivalent to any set of rows $r_1, \ldots, r_n$ such that $\forall r_i : W(r_i) = W(r)$, and $\bigcup_i L(r_i) = L(r)$. That is, r, $r_1, \ldots, r_n$ must have exactly the same Ws, and the combined Ls of $r_1, \ldots, r_n$ must form the same set as the Ls of r.

$C1$	$C2$	$C3$	$C4$
W	W	L	L

⇔

$C1$	$C2$	$C3$	$C4$
W	W	L	*e*
W	W	*e*	L

Proof: Suppose a ranking M puts on top of each L in r one of r's Ws. As any r_i has the same Ws, any L in any r_i will also be covered by a W under ranking M.

Conversely, suppose a ranking N is compatible with all rows $r_1, \ldots, r_n$. Consider some L of row r. Some r_i must have an L in the same constraint, and ranking N covers it with a W in one of $W(r_i)$. That W-constraint in r_i also has a W in r, by definition. Thus N covers the arbitrary L in r just as well.

Thus a ranking is compatible with r iff it is compatible with $r_1, \ldots r_n$. □

The remaining two pairs of operations are the non-trivial part of the set. Some OT rows may be superfluous in their tableaux: even if we delete them, the amount of information in the tableau does not change (e.g., the second row in Tableau (2) is superfluous.) By definition, subtraction or addition of such rows does not offend OT equivalence. What is non-trivial, though, is determining the exact formal conditions under which a row is superfluous. In the proof, I use the criterion by Prince (2002), featuring his operation of fusion.[7] One can provide an alternative characterization of superfluousness based on

[7] The operation of fusion on rows is defined by Prince (2002, page 8, Equation (12)). For tableau U, the fusion row fU has an e in the Ci cell iff all rows in U have an e in Ci; has an L iff some row in U has an L in Ci; and has a W otherwise, that is, when at least one row in U has a W in Ci, and all other rows have either Ws or es, but not Ls.

domination chains of constraints, but the proof based on such chains is more cumbersome.[8]

Using the fact that row order is not significant in a tableau, cf. Operation (3), we can safely use set notation for tableaux, understood as being parasitic on the notation for ERC sets: $T \setminus r$ denotes tableau T with row r subtracted; $T \cup U$ is a concatenation of tableaux T and U; and so forth.

(6) **Inference eliminations and introductions**: a row r entailed in tableau T by the rest of the tableau (that is, by $T \setminus r$) can be subtracted from T, or added back to tableau $T \setminus r$.

$C1$	$C2$	$C3$	$C4$
W	L	*e*	*e*
e	W	L	*e*
W	*e*	L	*e*

$\Longleftrightarrow$

$C1$	$C2$	$C3$	$C4$
W	L	*e*	*e*
e	W	L	*e*

Proof: Trivial. What is non-trivial is how to determine if r is entailed by $T \setminus r$. By Prop. 2.5 of (Prince 2002, p. 14), r is entailed by $T \setminus r$ iff there exists a subtableau U of $T \setminus r$ s.t. the fusion q (cf. Footnote (7)) of U entails r. In turn, q entails r either when r has no L-s and thus is compatible with any ranking, or when $W(q) \subseteq W(r)$ and $L(q) \supseteq L(r)$. □

[8] I provide the definitions of possible and maximal domination chains in (i), and the criterion of superfluousness based on them, without proof, in (ii):

(i) For a tableau T, a row $r_i \in T$, and a $Cj \in L(r_i)$, a **possible domination chain** is a sequence of constraints $\langle C_{k_1}, \ldots, C_{k_n} \rangle$ s.t. $C_{k_n} = Cj$, a single constraint never occurs twice in the chain, and for each C_{k_l}, $C_{k_{l+1}}$ there is a row $r_m \in T$ for which $C_{k_l} \in W(r_m)$, $C_{k_{l+1}} \in L(r_m)$. A **maximal possible domination chain** is a possible domination chain for which there is no $r_m \in T$ s.t. $C_{k_1} \in L(r_m)$.

(ii) **Superfluous row theorem**. A tableau $T = \langle r_1, \ldots, r_n \rangle$ entails a row q iff for each $Ci \in L(q)$, there exists such a row $r \in T$ in every maximal domination chain for Ci, r, and T, that there is a constraint C_{k_l} in it s.t. $C_{k_l} \in W(q)$.

Checking the criterion based on maximal chains does not require computing new rows, as the fusion criterion does. But it is easy to see from the cumbersomeness of the definitions that proving the criterion's correctness from first principles requires a bit of work. Therefore I simply reuse Prince's fusion-based result in the main text, referring the reader to Prince (2002) for proofs of its correctness.

To reduce the computational complexity of inference elimination, an RCD-based method is proposed by (Prince 2002, Sec. 5). Prince shows that instead of checking the fusions of all subtableaux, one may check whether $T \setminus r$ is consistent with the negative $\neg r$ of r, obtained by replacing all r's Ws with Ls and vice versa. For m rows, we need m such RCD-based checks. As Magri (2009) explains, RCD requires $m^2 n$ operations for a tableau with m rows and n constraints. The complexity of RCD-based inference elimination is thus polynomial, in contrast to subtableau-fusion version which is exponential in the number of constraints n.

Finally, not all Ws in an OT tableau are necessarily equal: there may be rows with "false Ws" such that there is no ranking compatible with the tableau which puts that W on top of any Ls in the row. As shown by (Prince 2006, p. 12), such false Ws may be replaced with an e without affecting the set of rankings the tableau is compatible with. An example of such a W is the W in the first row in $C3$ in the left tableau in Operation (7). The third row of the tableau necessitates ordering $C4$ over $C3$ in any compatible ranking M, and because of that the L in the first row may never be covered by the W in $C3$ in M. Therefore replacing that W with an e, as in the right tableau, does not offend OT-equivalence. The operation for doing such changes is called **Generalized Removal of W**, or GRW. We also introduce its inverse, **Generalized Introduction of W**, or GIW.

(7) **Generalized Removal of W (GRW) and Introduction of W (GIW)**: informally, a "false" W is a W whose replacement with an e does not change which rankings the tableau is compatible with. Thus a false W does not do any actual work. The example tableaux below may help visualize the phenomenon.

$C1$	$C2$	$C3$	$C4$
W	W	W	L
W	L	*e*	*e*
e	*e*	L	W

$\Leftrightarrow$

$C1$	$C2$	$C3$	$C4$
W	W	*e*	L
W	L	*e*	*e*
e	*e*	L	W

Turning to the formal definition: for rows r and r' such that instead of r's W in a fixed Ci, row r' has an e, consider a pair of T including r that is not entailed by the rest of T, and $T' := (T \setminus r) \cup r'$.

(That is, T' that is exactly like T, but with the W in Ci in row r replaced with an e.)

The claim is: T and T' thus defined are equivalent iff $T \setminus r$ entails the row q such that $W(q) := L(r) \cup (W(r) \setminus Ci)$, and $L(q) = \{Ci\}$.

Proof: Just as with inference eliminations, the fact that a false W can be replaced with an e is trivial. What is non-trivial is the criterion for false Ws: a W is false iff the row q as described above is entailed by $T \setminus r$. Prince (2006, p. 12) proves essentially that criterion in his (31) using fusion.[9] Yanovich (2011) provides a different proof in his (125) using partial OT rankings. The proof below is based on the idea of the proof in Yanovich (2011), but does not use either fusion or the apparatus of partial rankings.

Consider row r with a W in Ci, and row q defined as in the criterion above: $W(q)$ contains all W- and L-constraints of r except Ci, and the only L-constraint of q is Ci. We need to show that the W in Ci in r is false in tableau T precisely when the rest of the tableau, $T \setminus r$, entails the row q so constructed.

Without loss of generality, assume that r has only one L, in constraint Cj. (We have the right to assume that because we proved in 5 that any multiple-L row may be split into several single-L rows that are together equivalent to it.)

Suppose q is entailed by $T \setminus r$. We will prove that T is then equivalent to T', and thus the W Ci in r is false. Assume towards a contradiction that there is a ranking M which is compatible with T, but not with T'. That ranking M must be compatible with r, but not with r' which differs from it in that it has an e in Ci instead of a W. Then M must say that $Ci \gg Cj$ and that for every Ck from $W(r) \setminus Ci$, $Cj \gg Ck$: otherwise it would be compatible not only with r, but also with r'. But then M is incompatible with q: the L-constraint Ci dominates Cj in M, and then by transitivity any W-constraint Ck. This is contrary to assumption, and therefore there cannot be such an M. Furthermore, any ranking compatible with T' is bound to be compatible with T,

[9] Prince's theorem is slightly weaker compared to our formulation: Prince requires all rows in T to be not entailed by the rest of the tableau. His actual proof, though, only employs the fact that r is not entailed by $T \setminus r$, just as our proof does.

and thus we derive that if q is entailed by $T \setminus r$, then T and T' are OT-equivalent.

For the other direction, suppose q is not entailed by $T \setminus r$. We show that then there is a ranking compatible with T, but not T'. We need to show that there exists ranking M compatible with $T \setminus r$ that says $Ci \gg Cj \gg Ck$ for all $Ck \in W(r')$: such a ranking will be compatible with r, but not with r'. Towards a contradiction, suppose there is no such M. That is only possible if no ranking compatible with $T \setminus r$ says $Ci \gg Cj \gg Ck$. For the $Ci \gg Cj$ part, $T \setminus r$ cannot necessitate the opposite ordering $Cj \gg Ci$: if it did, then it would have entailed q, contrary to assumption. For $Cj \gg Ck$, suppose towards a contradiction that every ranking compatible with $T \setminus r$ says for some $Ck \in W(r')$ or other that $Ck \gg Cj$. That can only be if there is a row s in $T \setminus r$ with an L in Cj, and $W(s) \subseteq W(r')$. But if that is so, then the row s, and thus $T \setminus r$ as a whole, entail r: the L is in the same place in s and r, and $W(s) \subseteq W(r') \subset W(r)$. That is contrary to assumption, so if $T \setminus r$ does not entail either q or r, then there must be a ranking M compatible with $T \setminus r$ saying $Ci \gg Cj \gg Ck$ for all $Ck \in W(r')$. That M is compatible with T, but not with T', and thus witnesses that T and T' are not OT-equivalent: T is compatible with a larger number of rankings, thanks to the non-false W for which the criterion based on a specially constructed row q fails. □

We have now defined and proved correctness of five pairs of elementary operations preserving OT-equivalence of comparative tableaux. Those operations as such have been known before. What has not been known is that those five pairs form a *functionally complete* set: any transformation preserving OT-equivalence can be performed by applying a sequence of those elementary operations, as we will show in the next section.

The following easy-to-prove fact will become useful later:

(8) All operations in (3)–(7) have inverses: row swap is self-inverse; for the other four pairs, the two members of the pair are inverses.

What (8) means is that each sequence of applications of our elementary operations may be inverted: if we can derive from tableau T another tableau U using those operations, then we can also derive from U the original tableau T by applying the inverted form of the same sequence.

3 NORMAL FORM FOR OT TABLEAUX

In this section, we present core novel results of this paper: two theorems regarding the existence and uniqueness of normal form for comparative OT tableaux. Namely, we define a specific tableau format in Definition (9), and then prove that for each equivalence class $\mathscr{C}$, there exists exactly one tableau in such format, and moreover, that the normal form of a (finite) tableau is computable. Normal forms thus can serve as true representatives of their equivalence classes, giving us a handle on those.

It should be stressed that there is nothing particularly special about normal forms — in fact, as we will see in the next section, other forms may be proven to be usable as normal forms just as well. The reason we define the normal form in Definition (9) the way we do is simply that it is convenient for proof purposes. Nor is the form we chose new: Prince (2006, p. 6) defines essentially the same form in terms of ERCs, called the Minimal Primitive Generator, or MPG. Thus in this section we show that (the tableau counterpart of) an MPG is a true normal form for OT equivalence classes.

(9) **Normal form for OT tableaux**:

1. The only contradictory tableau in the normal form is the one-row tableau with a single L in the first constraint. We can refer to this special tableau as $T_\perp$.

2. Each row has at most a single L.[10]

3. There are no rows which can be inference-eliminated (see Operation (6)).

4. In multiple-W rows, there are no false Ws (see Operation (7)).

[10] Such single-L rows correspond to Primitive Ranking Conditions of Prince (2006).

5. The rows are ordered according to some strict total order of the set of all possible rows.[11] (For corresponding ERC sets, the notion of row order becomes irrelevant.)

Here is an example of a normal form tableau:

(10)

$C1$	$C2$	$C3$	$C4$	$C5$
W	*e*	L	*e*	*e*
W	*e*	*e*	L	*e*
W	*e*	*e*	*e*	L
e	W	*e*	L	*e*

Just calling something a normal form does not make it one. The results in (11)–(14) establish the fact that the class of tableaux defined in Definition (9) indeed has normal form properties.

(11) **Normal Form Existence Theorem**
An arbitrary (finite) tableau T can be transformed into an equivalent normal form tableau by a (finite) sequence of equivalence-preserving transformations in Operations (3)–(7).

(12) **Corollary to Theorem (11)**. Each non-empty equivalence class of tableaux contains at least one normal form tableau.

Proof of Theorem (11). We give an explicit procedure for transforming an arbitrary tableau so that it satisfies the requirements in Definition (9). For contradictory tableaux, we just add the row (L, *e*, *e*, ...), and subtract all others. If the tableau is not contradictory, we apply row splittings until all rows have at most one L (and are thus PRC-rows). Assuming the tableau is finite, we can eliminate all entailed rows by testing whether the fusions of subtableaux satisfy Prince's condition on entailment, see Operation (6). After that, we can similarly eliminate all false Ws from the resulting tableau by testing if the conditions for GRW, see Operation (7), are met (as all entailed rows were eliminated by that point, the row independence precondition of the criterion in Operation (7) is met). We finish the procedure by applying row swaps to get the ordering right. □

[11] The actual choice of ordering is irrelevant as long as all conceivable rows are strictly ordered. I will use the following: 1) let the first constraint where only one of r and q has a W be Ci; then the row with the W in Ci goes first; 2) for rows which have identical W-sets, the row which has an L in the first constraint where only one of them has an L goes first.

Proof of Corollary (12). Trivial: if there were no normal form tableau in a non-empty equivalence class, then Theorem (11) could not have been valid. □

Note that if the tableau is finite, the normalization procedure described in our proof of Theorem (11) is computable. This is important because if the normal form were not computable, we could not use it without restrictions in place of any other tableau in its equivalence class: we would not have been able to ensure we can actually derive one from the other in a finite amount of time. In fact, complexity analysis shows that normalization is not only computable, but quite efficient:

(13) Tableau normalization as defined in the proof of Theorem (11) runs in time polynomial in the number of rows m and the number of constraints n.

Proof of Theorem (13). Consider tableau T with m rows and n constraints. Consistency check may be performed through fusing all subtableaux of T and checking if any resulting fused row has only Ls — or equivalently and faster using RCD, as shown by Prince (2002, Section 4). To perform RCD, we need m^3n operations (Magri 2009, p. 371). Next, we do row splittings, which for any of the m rows cannot result in creating more than n rows of n constraints each, so this requires at most mn^2 operations. The number of rows in the resulting split tableau is not greater than mn. Next, we check for entailed rows to eliminate. As we discussed above regarding Operation (6), rather than doing subtableau fusion, exponential in the number of rows m, we can do instead m RCD-based checks as described by Prince (2002, Section 5). We have mn rows, and each RCD involves $(mn)^3n$ operations, so overall we need m^3n^4 operations for this step. Finally, we need to check for false Ws. For that we check every W in mn rows, so at most this would be mn^2 checks (actually, much less, as the same row cannot contain both n Ls and n Ws, but we can ignore this.) Each test involves checking whether the rest of the tableau entails a specially constructed row for each particular W. Again, the cost of an entailment check for a single row and a tableau with mn rows is m^3n^4, so overall we have at most m^4n^6 operations. This will be the dominating term in our complexity estimate. The time complexity of normalization is thus polynomial, which is very good. □

From Theorems (11) and (12), we know that each equivalence class has at least one normal form tableau. But can a class contain more than one normal form? Theorem (14) shows that it cannot, and thus a normal form tableau *defines* its class: it is its unique representative. To prove that fact, we will need to use relatively complex ranking-construction techniques.

(14) **Normal Form Uniqueness Theorem**
In each equivalence class of OT tableaux, there is at most one normal form tableau.

Proof of Theorem (14). We show that any two distinct normal form tableaux T and U belong to different equivalence classes.

Pick some row r from T which is not shared by U (in case $T \subset U$, we immediately derive the conclusion by considering a row from U that is not in T, and the fact that T cannot entail that row). Either our pick r is entailed by U, or it is not. In case r is not entailed by U, there is some ranking M compatible with U, but not with r, and thus not with T, so U and T are not OT-equivalent.

The interesting case is when U entails the row r we picked. We will show that in that case, there must be some ranking compatible with T, but not with U. We pick a minimal subtableau V of U that still entails r. As V entails r, every ranking compatible with V must also put one of r's Ws on top of r's L. That can only be if there is a row $q \in V$ which has an L in the same constraint where r has an L. Let's call that constraint Ci.

Suppose towards contradiction that T and U are equivalent, that is, compatible with exactly the same rankings. Consider some M compatible with T, and accounting for V "in the minimal possible manner": let M contain the domination chain $Ck_1 \gg Ck_2 \gg \ldots \gg Ci$ where each pairwise ranking $Ck_1 \gg Ck_2$, ..., $Ck_n \gg Ci$ accounts for one of the rows in V, but no other pairwise rankings accounting for any of V's rows. As V is in normal form and all its Ws are not false, it must be possible to construct such an M. As V entails r, constraint Ck_1 is a W-constraint in r.

$T \setminus r$ cannot entail V: if it did, it would have entailed r by transitivity, which is contrary to the normal form assumption. Therefore it must be possible to lower one of Ck_i constraints below Ci building a ranking M' which is still compatible with $T \setminus r$, but not with V. As V entails r by assumption, M' must also be incompatible with r. But that can only be if the lowered constraint has to be Ck_1, a W-constraint in r, for otherwise M' would have still said $Ck_1 \gg Ci$.

We modify M' as follows: raise Ck_1 just on top of Ci, but below Ck_n, resulting in $M'' = \ldots Ck_2 \gg Ck_3 \gg \ldots \gg Ck_n \gg Ck_1 \gg Ci$. That ranking M'' is incompatible with V, because by construction there must have been a row in V for which we needed the pairwise ranking $Ck_1 \gg Ck_2$, and M'' says $Ck_2 \gg Ck_1$. But at the same time M'' is compatible with r, as it puts one of its Ws on top of its L. Now compare M' and M'', and consider their compatibility with $T \setminus r$. M' was compatible with $T \setminus r$. M'' differs from it in that it says $Ck_1 \gg Ci$ instead of $Ci \gg Ck_1$. That change could not make M'' incompatible with $T \setminus r$: the initial ranking M also said $Ck_1 \gg Ci$ and was compatible with $T \setminus r$. Therefore we have built a ranking, namely M'', which is compatible with $T \setminus r$ and with r, but not with V. This ranking witnesses that T and U are not equivalent. □

Theorems (11) and (14) together entail that there is exactly one normal form tableau per equivalence class. Thus a tableau as described in Definition (9) is a true normal form: a full-fledged representative, or a "name", of its equivalence class.

In practical terms, that means that in our proofs, we can capitalize on the many nice properties of normal forms, knowing that the results will generalize to arbitrary tableaux. In the next section, we illustrate that the use of the normal form results in several simple corollaries.

4 CAPITALIZING ON THE NORMAL FORM RESULTS

Theorems in (15), (16) and (18) serve two purposes. First, they have independent value, especially the proof that Brasoveanu and Prince's SKB bases are unique in their equivalence classes. Second, the proofs of these statements illustrate how one can use the normal form results in practice to handle equivalence classes of OT tableaux.

(15) Operations (3)–(7) form a *functionally complete set*: any tableau can be transformed into any equivalent tableau by a sequence of such operations.

Proof of (15). By Theorems (11) and (14), any pair of equivalent tableaux may be transformed into the same normal form tableau by a sequence of operations in Operations (3)–(7). To conclude the proof, we observe that an inverted sequence transforms the normal form back into the original tableau. By normalizing the first tableau, and then denormalizing it by applying the inverted sequence built for the second tableau, we transform the first tableau into the second. □

(16) Equivalence of finite OT tableaux is computable in polynomial time.

Proof of (16). To test tableaux T and U for equivalence, it suffices to normalize both and check whether the resulting normal forms are the same. All operations are computable, for finite tableaux.

The complexity of this test is polynomial: by Theorem (13), the time complexity of normalization is polynomial in the number of rows m and the number of constraints n, and we need two such normalizations, plus a comparison of two resulting normal form tableaux which is also polynomial in m and n for the original tableaux. This fairly moderate complexity may be compared with the enormous factorial complexity of the brute-force test for equivalence that involves testing every possible ranking for compatibility with each tableau, cf. Footnote 2. □

Brasoveanu and Prince (2011) define a dense format of tableaux called the Skeletal Basis (SKB) and an algorithm turning an arbitrary tableau into an equivalent tableau in that format. An SKB of tableau T is a tableau T' such that 1) there is no OT-equivalent tableau with a smaller number of rows; and 2) no other equivalent tableau of the same cardinality has more *e*s. Tableau (17) is the Skeletal Basis of the normal form tableau in Tableau (10):

(17)

$C1$	$C2$	$C3$	$C4$	$C5$
W	*e*	L	L	L
e	W	*e*	L	*e*

Brasoveanu and Prince (2011) claim to have proven, in an unpublished manuscript, the fact that *for a single tableau*, the SKB basis is unique. (Prince (2006, page 6) derives from that the result that MPGs, corresponding to our normal forms, are also unique for a single tableau.) Using our uniqueness theorem for normal forms in Theorem (14), we prove a much stronger result for SKBs in Theorem (18): each *equivalence class of tableaux* has a unique SKB.

(18) Each equivalence class of OT tableaux has exactly one tableau in the Skeletal Basis (SKB) form of Brasoveanu and Prince (2011).

Proof of (18). By showing that SKBs are in one-one correspondence with normal forms.

If we apply all possible row mergers to a normal form tableau, we get an SKB: the original normal form tableau did not have superfluous rows, so the quantity of the rows in the resulting tableau will be minimal; furthermore, as the normal form tableau does not contain any false Ws, the resulting tableau will have the maximal number of *e*s.

In the other direction, if we split all rows of an SKB into one-L rows, there can be no superfluous rows in the result (otherwise the L corresponding to a superfluous row could have been replaced with an *e*, contrary to the definition of an SKB which must have as many *e*s as possible); as for false Ws, there can be none in the SKB tableau itself, and after all row splittings are applied, no new false Ws can arise (if a false W could arise in one of the resulting one-L rows, then the same W would have been false even before splitting).

What remains is to show that there can be no two SKBs in the same equivalence class. Suppose towards contradiction there are two SKBs S_1 and S_2. They both normalize to the same normal form tableau by the procedure above. From the definition of SKB, only row splittings are required. Pick an arbitrary set of rows $r_1, \ldots r_n$ with Ws in the same constraints from the resulting normal form tableau. If S_1 and S_2 each have only one row splitting into this same set, that must be the same row. If S_1 and S_2 have more than one row splitting into this set, we can actually merge those rows into just one, resulting in a smaller equivalent tableau S_3, contrary to the assumption of S_1 and S_2's minimality. Thus either $S_1 = S_2$, or they are not minimal possible size in their equivalence class. Therefore there is only one SKB per class. □

Theorem (18) essentially means that all useful results about normal forms may be transferred to SKBs. For instance, the equivalence test in Theorem (16) may be replaced by an equivalence test comparing SKBs derived using Brasoveanu and Prince's Fusional Reduction algorithm. With 18 in hand, we may employ Brasoveanu and Prince's SKBs as representatives of their equivalence classes instead of our normal forms. Normal forms are often more convenient in complex proofs, because the relations between constraints in them are maximally untangled; but SKBs are more useful when it becomes convenient to have smaller-sized representatives.

5 CONCLUSION

We defined a normal form for OT tableaux, and showed that there is exactly one normal form in each equivalence class of OT tableaux. Moreover, we have demonstrated that each OT tableau can be computably normalized by a sequence of five pairs of previously known equivalence-preserving transformations in Operations (3)–(7). The computational cost of normalization is only polynomial in the number of rows m and constraints n, thanks to the use of the efficient RCD-based algorithm for entailment checking proposed by Prince (2002, Section 5).

Those results provide us with a handle on equivalence classes of OT tableaux: using them, we may reason about tableaux without any loss of generality while only considering normal forms. The examples in Section 4, including Theorem (18) stating that Brasoveanu and Prince's Skeletal Bases are unique in their equivalence classes, illustrate how to capitalize on the presented OT normal form theorems.

REFERENCES

Adrian Brasoveanu and Alan Prince (2005), Ranking and Necessity, Part I, Rutgers Optimality Archive 794.

Adrian Brasoveanu and Alan Prince (2011), Ranking and Necessity: the Fusional Reduction Algorithm, *Natural Language and Linguistic Theory*, 29(1):3–70, *revised version of Brasoveanu and Prince (2005)*.

Bruce HAYES (1997), Four Rules of Inference for Ranking Argumentation, ms., UCLA. http://www.linguistics.ucla.edu/people/hayes/otsoft/argument.pdf.

Renè KAGER (1999), *Optimality Theory*, Cambridge University Press, Cambridge.

Giorgio MAGRI (2009), *A Theory of Individual-Level Predicates Based on Blind Mandatory Implicatures. Constraint Promotion for Optimality Theory*, Ph.D. thesis, MIT. http://dspace.mit.edu/handle/1721.1/55182.

Alan PRINCE (2000), Comparative tableaux, Rutgers Optimality Archive 376.

Alan PRINCE (2002), Entailed Ranking Arguments, Rutgers Optimality Archive 500. http://roa.rutgers.edu/article/view/510.

Alan PRINCE (2006), No more than Necessary: beyond the Four Rules, and a bug report, Rutgers Optimality Archive 882. http://roa.rutgers.edu/article/view/905.

Alan PRINCE and Paul SMOLENSKY (1993), Optimality Theory: Constraint Interaction in Generative Grammar, Rutgers Optimality Archive 537. http://roa.rutgers.edu/article/view/547.

Alan PRINCE and Paul SMOLENSKY (2004), *Optimality Theory: Constraint Interaction in Generative Grammar*, Blackwell, Oxford.

Igor YANOVICH (2011), On sets of OT rankings, Rutgers Optimality Archive 1149. http://roa.rutgers.edu/article/view/1203.