Categorial Parsing and Normalisation

Hans Leiß
(editor)

DYANA-2

Dynamic Interpretation of Natural Language
ESPRIT Basic Research Project 6852

Deliverable R1.1.A
August 1993

Categorial Parsing and Normalisation

Hans Leiß
Universität München
(editor)

authors:
Herman Hendriks
Johan van Benthem & Grisha Mints
Martin Emms & Hans Leiß
Wojciech Buszkowski
Yde Venema
Dirk Roorda

DYANA-2

Dynamic Interpretation of Natural Language
ESPRIT Basic Research Project 6852

Deliverable R1.1.A
August 1993

ESPRIT Basic Research Action 6852 DYANA-2

Dynamic Interpretation of Natural Language

Universiteit van Amsterdam
Institute for Logic, Language and Computation (ILLC)

University of Edinburgh
Centre for Cognitive Science (ECCS)

Universität München
Centrum für Informations- und Sprachforschung (CIS)

Universitetet i Oslo
Department of Linguistics and Philosophy (ILF)

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung (IMS)

Universität Tübingen
Seminar für Sprachwissenschaft (SfS)

Rijksuniversiteit Utrecht
Research Institute for Language and Speech (OTS)

For copies of reports, updates on project activities and other DYANA-related information, contact:

The DYANA-2 Project Administrator
ILLC/Department of Philosophy
University of Amsterdam
Nieuwe Doelenstraat 15
NL-1012 CP Amsterdam

Contents

Introduction by the editor v

Task 1.1, subtask 1: Categorial Parsing and Normalisation
Lambek Semantics
Herman Hendriks 1

Comments on Hendriks' "Lambek Semantics"
Johan van Benthem and Grisha Mints 73

Additional Contribution
The Cut-Elimination Theorem for the Second Order Lambek Calculus
Martin Emms and Hans Leiß 77

Comments on "The Cut-Elimination Theorem for the Second Order Lambek Calculus"
Wojciech Buszkowski 101

Additional Contribution
Meeting a Modality? Restricted Permutation for the Lambek Calculus
Yde Venema 105

Comment on Yde Venema
Dirk Roorda 123

Introduction

This collection contains three papers in Dyana-2's Area 1 on "Grammar architecture", all contributing to Task 1.1 on "Categorial grammar and type theory".

The first paper by Hendriks is devoted to categorial parsing and normalization, (Subtask 1). Its main contribution is a variant L^* of Lambek's calculus L (without product) that solves the "spurious ambiguity" problem, i.e. ensures that each sequent (expressing a subsumption judgement between a combination of types and a type) has only as many proofs as it has meanings. This allows to identify meanings with proof structure, and eliminates unnecessary proof search during categorial parsing.

The two other papers of this collection are concerned with two extensions of Lambek's calculus by additional type constructors.

The one by Venema studies a type constructor ∇ used to categorize as "special" those expressions of a category X which can be moved away from their expected positions. On the one hand, it is investigating how to relax structure sensitive type deduction by "structural modalities" in a principled way. On the other hand, however, it shows that –for the particular operator in question– the modal interpretation is not primitive, but can be decomposed using a base type of commutable elements, reflecting rather directly the intuitive meaning of movability. It would be interesting to see similarly direct interpretations of other structural modalities.

The other paper, by Emms and Leiß, is concerned with type abstraction as a means to capture type uniformities in lexical assignments. A "polymorphic" extension of Lambek's calculus by type quantifiers is shown to have the cut-elimination property. This has to be seen as a first glance only at mathematical properties of a polymorphic version of Lambek's calculus; basic questions like whether provability is decidable or whether there is a natural semantics (useful for establishing unprovability claims) remain open. In particular, the effects of such an extension to categorical parsing are not quite understood.

It may be useful to point out the role of these papers for the goals of the Dyana-2 project.

Lambek Semantics

The main paper of the collection is the one by Herman Hendriks on 'Lambek Semantics'. Of course, Lambek semantics is based on the Curry-Howard correspondence between proofs in Natural Deduction (ND) systems and terms in typed λ-calculus: the meaning of a proof is just the function denoted by the λ-term that fairly directly corresponds to the proof. However, when moving from natural deduction to Gentzen-style sequent systems –the preferred proof system in Lambek calculus–, several proofs in the sequent calculus may translate into the same ND-proof or λ-term. Thus, different proofs do not necessarily reflect different readings of the same syntactic judgement. Restriction to "normal" sequent proofs is needed to avoid this "spurious" kind of ambiguity of meaning of sequents.

Hendriks' paper contains (i) a comparison between various approaches to proof normalisation for the Lambek calculus (without product), (ii) a new proposal what a 'normal' proof in Lambek calculus should be, which is shown to solve the 'spurious ambiguity' problem, and (iii) a demonstration that Lambek's calculus with several atomic categories can be embedded in Lambek's calculus with a single atomic category by interpreting several atoms by suitable complex categories.

The relation of the results obtained to the logical background is sketched in the Comment by J. van Benthem and G. Mints. To the non-specialist, some further comments concerning the content of the paper may be helpful.

First, a modification in the λ-term associated to a sequent proof is motivated, and it is shown that Lambek's cut-elimination procedure essentially preserves the term associated to a proof: as one expects, the key case of cut-elimination amounts to a β-reduction in the corresponding proof terms. Thus λ-terms associated to cut-free Lambek proofs are in β-normal form.

Since it is possible that different cut-free Lambek-proofs are annotated by different β-normal form terms *with the same denotation*, some of these proofs have to be banned. This is what is achieved using the modified calculus L^*, whose proofs correspond to λ-terms in $\beta\overline{\eta}$-normal form. Since semantically equivalent $\beta\overline{\eta}$-normal forms are identical, each L^*-provable sequent has a unique proof. It is then possible to solve the spurious ambiguity problem by translating L^*-proofs into L-proofs with certain properties (and vice versa), giving unique "normal" proofs for the system L.

Next, a comparison is made between this and two other proposals to deal with spurious ambiguity. Moortgat (1990) suggested to solve this problem using "partial deduction", i.e. by a precompilation of the lexical type assignment into axioms and rules that replace those of L in the proof search. Hendriks gives a formal elaboration of this proposal and, using a translation of certain proofs in his L^*-calculus to partial-deduction-proofs, shows that it indeed solves the problem of spurious ambiguity when restricted to sequents with atomic goal categories. On the other hand, Roorda (1991) suggested to adapt Girard's notion of proof net in linear logic to Lambek calculus. Spurious ambiguity was hoped to be eliminated this way since proof nets abstract from various orderings of rule applciations in the sequent calculus. However, as Hendriks shows, this is not sufficient to provide unique normal proofs: by relating proof nets to L^* proofs, it is revealed that several nets correspond to the same L^*-proof, and hence represent the same meaning.

These results raise several questions:

1. As is pointed out in Girard e.a., the Curry-Howard correspondence between λ-terms and Natural Deduction proofs works best for the connectives of conjuction, implication and universal quantification: since these are the connectives of Lambek calculus (with its extensions by unification and polymorphic types), it may indeed –as the Comment by van Benthem and Mints points out– be better to use natural deduction instead of sequent calculus in studying categorial grammar, getting rid of the spurious

ambiguity problem right from the beginning.

2. One would like to understand precisely the effect of the proof normalzation via L^* on categorial parsing: is it possible – for example by studying the translations between the proof systems given– to find bounds for the differences in (time) complexity of proof search with respect to L, L^*, partial deduction proofs and proof nets?

Movement: Lambek calculus with restricted permutation

The paper by Venema contributes to the aim of relaxing structure sensitive type deduction by allowing controlled use of 'structural' proof rules in an extension of Lambek's calculus. As is familair from linear logic, "structural modalities", i.e. modal connectives in the type language, are used to mark types that may be used more freely in a derivation than others.

A particular case is Lambek calculus extended by a modal operator which allows a limited use of the exchange rule. This is useful linguistically when categorizing expressions in which some constituents are moved from their expected position. While it is clear what the right proof rules for a structural modality for movability are, it is less clear why the interpretation should be a modal one, i.e. based on frame structures used to interpret modal operators. Indeed, earlier proposals by Hepple and Morill suggest to interpret Lambek calculus with an exchange modality in monoids with a distinguished submonoid of commutable elements.

The paper by Venema gives a reformulation of Lambek calculus with an exchange modality, but reflects in a direct and appealing way the intuitive notion of a monoid with distinguished submonoid of commutable elements. Cut-elimination, dedcidability, and completeness of the calculus with respect to this interpretation are given.

From a linguistic point of view, one would like to know whether refined notions of movement can also be captured: for example, could one describe movement to a certain position, or to everywhere except a certain position? A comparison with treatments of movement phenomena in other grammar models would certainly be helpful in understanding the potential of the proposed system.

Polymorphism: Lambek Calculus with strong lexical type uniformity

The main motivation for the introduction of (universal) type quantifiers – sometimes called inherent polymorphism– is to capture uniformities in the set of possible types of an expression. If infinitely many types of an expression fall under a common schema, it is most natural to extend the type language by schematic types. This approach has been very fruitful in the study of programming languages; of course, the relevance of such strong assumptions about type uniformity in natural language needs independent linguistic motivation.

While some linguistic examples are given in another deliverable of Dyana-2, the paper by Emms and Leiß begins a systematic investigation of the mathematical aspects of polymorphic Lambek calculus. The current, apparently rather strong system of second order Lambek calculus is shown to satisfy the cut-elimination property. A demonstration of the decidability of provability (needed for categorial parsing with such a system) is missing still, and perhaps impossible to give. Hence it may well be that the system used in this paper needs restrictions motivated by both linguistic and computational considerations – even if some mathematical questions are best studied in a "pure" version.

From a linguistic point of view, universal quantification over arbitrary types may be too strong and may have to be restricted to quantification over a definable subset of types –for example, the conjoinable types–, or to bounded quantification when a notion of subtyping is present. Also, there are questions of interplay between type quantification and derivational polymorphism.

It seems possible to extend the 'Lambek semantics' of derivations from the simple to the second order Lambek calculus, using terms of the second-order λ calculus. However, a clear picture of the semantics of quantified types in the first place is needed, and an elaboration of the connection to various aspects of second order λ-calculus in programming.

The three papers in this collection are dealing with mathematical aspects of the formalisms used in the grammar architecture area of Dyana-2. It is hoped that all three papers bring out their linguistic relevance more clearly in published and more complete versions.

München, July 1993
Hans Leiß

Task 1.1, subtask 1

Categorial Parsing and Normalisation

Lambek Semantics

Herman Hendriks

(Utrecht University)

Comments on Hendriks' "Lambek Semantics"

Johan van Benthem and Grisha Mints

(University of Amsterdam & Stanford University)

The Cut-Elimination Theorem for the Second Order Lambek Calculus

Martin Emms and Hans Leiß
(Munich University)

Comments on the paper "The Cut-Elimination Theorem for the Second Order Lambek Calculus" by Martin Emms and Hans Leiss

Wojciech Buszkowski
(Adam Mickiewicz University, Poznan)

Meeting a Modality?
Restricted Permutation for the Lambek Calculus

Yde Venema

(Utrecht University)

Comment on Yde Venema

Dirk Roorda

(Groningen University)