The Lost Combinator

Citation for published version:

Digital Object Identifier (DOI):
10.1162/coli_a_00328

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Computational Linguistics

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Let me begin by thanking the Association for Computational Linguistics and its Executive Committee for conferring on me the great honor of their lifetime achievement award for 2018, which of course I share with all the wonderful students and colleagues that have made many essential contributions to this work over many years.

At the heart of the work that I have been pursuing over my research lifetime so far, whether in parsing and sentence processing, spoken language understanding, semantics, or even in musical understanding by machine, there lies a theory of natural language grammar that brings parsing, compositional semantics, statistical modeling and logical inference into the closest possible relation. This theory of grammar is combinatory, in the sense that its operations are type-dependent and restricted to strictly string-adjacent phonologically or graphologically-realized inputs, and categorial, in the sense that those operands pair a syntactic type with a type-transparent semantic representation or logical form.

I'd like to use this opportunity to briefly address three questions that revolve around the theory of grammar, both combinatory and otherwise. The first question concerns the way that Combinatory Categorial Grammar (CCG) was developed with a number of colleagues, over a number of stages and in slightly different forms. The second is an essentially evolutionary question of why natural language grammar should take a combinatory form. The third question is that of what the future holds for CCG and other structural theories of grammar in Computational Linguistics and NLP in the Age of Deep Learning.

I have called this talk “The Lost Combinator” in homage to the Victorian era poem “The Lost Chord”, in the hope of suggesting that the theoretical development of CCG has always been
empirical, rather than axiomatic, in search of the simplest explanation of the facts of language, rather than for confirmation of linguistic received opinion, however intuitively salient.

1. In the Beginning

In the late ’60s (when I was a psychology undergraduate at the University of Sussex under Stuart Sutherland, and then started as a graduate student in artificial intelligence at Edinburgh under Christopher Longuet-Higgins), a broad community of theoretical linguists, psychologists, and computational linguists saw themselves as all working on the same problem, under the definition provided by the “transformational” theory of grammar proposed by Chomsky (1957, 1965), using theories of psycholinguistic processing, language acquisition, and language evolution proposed by Lashley (1951), Miller, Galanter, and Pribram (1960), Miller (1967), and Lenneberg (1967), theories of natural language semantics proposed by Carnap (1956), Montague (1970), and Lewis (1970), and computational models of parsing such as those proposed by Thorne, Bratley, and Dewar (1968), and Woods (1970). (I myself was so convinced that this program would succeed that I believed it was time to apply the same methods to other cognitive faculties, taking as my research project for PhD their application to the interpretation of music by machine, following the lead of Max Clowes (1971) in machine vision.)

Almost immediately, this consensus fell apart. First, Chomsky himself was among the first (1965) to recognize that transformational rules, though descriptively revealing, were so expressive as to have little explanatory force, and required many apparently arbitrary constraints (Ross, 1967). Second, psychologists realized that psycholinguistic measures of processing difficulty of sentences bore almost no relation to their transformational derivational complexity (Fodor, Bever, and Garrett, 1974; Marslen-Wilson, 1973) Finally, computational linguists attempting to implement transformational grammars as parsers realized that they were spending all their time implementing even more constraints on rules, in order to limit search arising from overgeneration (Friedman, 1971; Gross, 1978). (Meanwhile, I realized that the problem had not in fact been solved, and returned to natural language processing, thanks to a postdoc at Sussex with Philip Johnson-Laird.)

This disillusion wasn’t just a case of internal academic squabbling. There were also a couple of influential reports commissioned by the US and UK governments that ended funding for Machine Translation (MT) and Artificial Intelligence (AI) (Pierce et al., 1966; Lighthill, 1973). As a result of the second of these reports, which determined that AI was never going to work, PhDs in artificial intelligence like my classmate Geoff Hinton and myself spent ten years or so after graduation in psychology departments (in my case, at the Universities of Sussex and
Warwick), until yet another report said AI was working after all and that Britain and the US were falling behind Japan in this vital area. As a result, I could get hired again in CS, first briefly back at Edinburgh, and then at the University of Pennsylvania (I learned a lesson from this odyssey that I have tried to remember whenever I have been appointed to a committee to report on anything, which is that while reports very rarely do any good, they can very easily do a great deal of harm.)

Meanwhile, as a result of these conflicts, the scientific study of language fragmented. The linguists swiftly abjured any responsibility for their grammars (“Competence”) bearing any relation to processing (“Performance”). Since the psychologists could hardly abandon Performance, they in turn became agnostic about grammar, retreating to context-free surface grammar (which they tended to refer to as “parsing strategies”), or a touchingly optimistic belief in its emergence from neural models. Meanwhile, the computational linguists (whose machines were growing exponentially in size and speed from the 16K byte core of the machine that supported the whole group when I started my graduate studies, on to levels that would soon permit parsing the entire contents of the then embryonic web) similarly found that very little of what the linguists and psychologists cared about was usable at scale, and that none of it significantly improved overall performance over very much simpler context-free or even finite-state methods that the linguists had shown to be incomplete. The reason of course was Zipf’s law, which means that the events with respect to which the low-level methods are incomplete are off in the long tail.

It also became apparent to a few computationalists working on speech, MT, and IR that the real problem was not grammar but ambiguity and its resolution by world-knowledge, and that the solution lay in probabilistic models (Bar-Hillel, 1960/1964; Spärck Jones, 1964/1986; Wilks, 1975; Jelinek and Lafferty, 1991) (although it was not immediately apparent how to combine statistical models with grammar-based systems without making obviously false independence assumptions).

Nevertheless, as any red-blooded psychologist had always insisted, the divorce between competence and performance that everyone else had accepted didn’t make any sense. The grammar and the processor had to have evolved in lock-step, as a package deal, for what could be the evolutionary selective advantage of a grammar which you can’t process, or a parser without a grammar?

It seemed equally obvious that surface syntax and the underlying semantic or conceptual representation must also be closely related, since the only reasonable basis for child language acquisition that has ever been on offer is that the child attaches language-specific grammar to a universal conceptual relation or “language of mind” (Miller, 1967; Bowerman, 1973; Wexler and Culicover, 1980). It seemed to follow that radically new theories of grammar were needed.
2. The Problem of Discontinuity

Theoretical linguists agree that the central problem for the theory of grammar is discontinuity or non-adjacent dependency between predicates and their arguments:

(1)  

Chomsky described discontinuity in terms of movement, which was known to be formally very unconstrained. By contrast, the ATN parser used in the LUNAR project (Woods, Kaplan, and Nash-Webber, 1972) reduced all discontinuity to local operations on registers (Thorne, Bratley, and Dewar, 1968; Bobrow and Fraser, 1969; Woods, 1970)

In particular, unbounded \textit{wh}-dependencies like the above were handled by: (a) putting a pointer into a * or HOLD register as soon as the “which” was encountered without regard to where it would end up, and; (b) retrieving the pointer from HOLD when the verb needing an object “had” was encountered without regard to where it had started out. (It also included an ingenious mechanism for coordination called SYSCONJ that one finds even now being reinvented on an almost yearly basis—cf. Woods, 2010.) A * register was also used for \textit{wh}-constructions within a systemic grammar framework by Winograd (1972: 52-53) in his inspiring conversational program SHRDLU.

However, it was unclear how to generalize the HOLD register to handle the multiple long-range dependencies, including crossing dependencies, that are found in many other languages. In particular, if the HOLD register were assumed to be a stack, then the ATN becomes a two-stack machine (since we are already implicitly using one stack as a PDA to parse the context-free core grammar).

On the computational side at least, the reaction to this impass took two distinct forms. Both reactions took the form of trying to reduce the two major operators of the transformation theory, substitution of immediate constituents or what is nowadays called “Merge”, and “Move” or displacement of non-immediate constituents, to one. On the one hand, Lexical Functional Grammar (LFG, Bresnan and Kaplan, 1982) and Head-driven Phrase Structure Grammar (HPSG, Pollard and Sag, 1994) followed Kay (1979) in making unification the basis of both merge and move. Since unification can pass information across unbounded structures, this can be thought of as reducing Merge to Move.
On the other hand, Generalized Phrase Structure Grammar (GPSG, Gazdar, 1981), Tree Adjoining Grammar (TAG Joshi and Levy, 1982), and Combinatory Categorial Grammar (CCG, Ades and Steedman, 1982), sought to reduce Move to various forms of local Merger. In particular, the latter authors suggested that the same stack could be used to capture both long-range dependency and recursion in CCG.\(^1\)

3. Why Does Natural Language Allow Discontinuity?

Natural language grammar exhibits discontinuity because semantically language is an applicative system. Applicative systems (such as programming languages) support the twin notions of: (a) Application of a function/concept to an argument/entity; and, (b) Abstraction, or the definition of a new function/concept in terms of existing ones.

Language is in that sense inherently computational. It seems to follow that linguistics is (or should be) inherently computational as well. (Of course, it does not follow that computationalists have nothing to learn from linguistics.)

There are two ways of modeling abstraction in applicative systems: (a) Taking abstraction itself as a primitive operation (\(\lambda\)-calculus, LISP):

\[
\begin{align*}
\text{(2) } & \quad \text{a. } \text{father Esau } \Rightarrow \text{ Isaac} \\
& \quad \text{b. } \text{grandfather } = \lambda x. \text{father}(\text{father } x) \\
& \quad \text{c. } \text{grandfather Esau } \Rightarrow \text{ Abraham}
\end{align*}
\]

(b) Defining abstraction in terms of a collection of operators on strictly adjacent terms aka Combinators, such as function composition (Combinatory Calculus, MIRANDA).

\[
\begin{align*}
\text{(3) } & \quad \text{b'. } \text{grandfather } = B/\text{father father}
\end{align*}
\]

The latter does the work of the lambda calculus without using any variables.

Despite the resemblance of the “traces” (or copies) and “operators” (or complementizer positions) of the transformational theory to the \(\lambda\)-operators and variables of applicative systems of the first kind, natural language actually seems to be a system of the second, combinator kind. The evidence stems from the fact that natural language deals with all sorts of fragments that linguists don’t normally think of as semantically typable constituents, without the use of any phonologically realized equivalent of variables, such as pronouns:

\[\text{father Esau } \Rightarrow \text{ Isaac} \]

---

\(^1\) As well as at Warwick and Edinburgh, much of the early work on CCG was done during 1980/81 as a visiting fellow at the University of Texas at Austin, under funding from the Sloan Foundation to Stanley Peters and Phil Gough, my first introduction to the United States.
(4)  
a. Give [Anna books]?
    and [Manny records]?

b. (Mother to child): There’s a DOGGIE! [You LIKE]?
    # the doggie.

c. Food that you must [wash VP/NP [before eating](VP,VP)/NP].

d. ik denk dat ik1 Henk2 Cecilia3 [zag1 leren2 zingen3]?

These fragments are diagnostic of a Combinatory Calculus based on B^n, T, and the “duplicator” Sn, plus application (Szabolcsi, 1989; Steedman, 1987; Steedman and Baldridge, 2011).^2

4. Combinatory Categorial Grammar (CCG)

CCG lexicalizes all bounded dependencies, such as passive, raising, control, exceptional case-marking, etc., via lexical logical form. All syntactic rules are Combinatory—that is, binary operators over contiguous phonologically realized categories and their logical forms. These rules are restricted by a Combinatory Projection Principle (CPP) which in essence says they cannot override the decisions already taken in the language-specific lexicon, but must be consistent with and project unchanged the directionality specified there. All such language-specific information is specified in the lexicon: the combinatory rules like composition are free and universal. All arguments, such as subjects and objects, are lexically type-raised to be functions over the predicate, as if they were morphologically cased as in Latin, exchanging the roles of predicate and argument.

All long range dependencies are established by contiguous reduction of a wh-element such as (N\N)/(S/NP) with an adjacent non-standard constituent with category S/NP, formed by rules of function composition.

\[
\text{company that Verizon owns}
\]
\[
N \quad (N\N)/(S/NP) \quad S/(S/NP_{3s}) \quad (S/NP_{3s})/NP
\]
\[
\text{owns}
\]
\[
\rightarrow
\]
\[
B
\]
\[
\text{companyx} \land \text{ownsxverizon}
\]

---

^2 The combinators are identified as B^n, T, and Sn for historical reasons. This combinatorial calculus is distinct from the categorial type-logic stemming from the work of Lambek (1958) that similarly underpins Type-Logical Grammar (Oehrle, 1988; Hepple, 1990; Morrill, 1994; Moortgat, 1997), in which the grammar is a logic, rather than a calculus, and some but not all of the CCG combinatory rules are theorems.
The combinatory rules synchronize composition of the syntactic types shown above with corresponding composition of logical forms (suppressed in the derivations above), to yield the logical forms shown as \( \lambda \)-terms for the resulting nouns \( N \).

To capture the construction in (4c), whose syntactic derivation we pass over here, we also need rules based on the duplicator \( S \):

$$
\text{(7) a. "wash X before eating X"}
$$

$$
\text{b. } VP/NP : \lambda x. \text{before}(\text{eat}x) (\text{wash}x) \equiv S(B\text{before eat})\text{wash}
$$

To capture constructions like (4d) (whose syntactic derivation is similarly suppressed), we also need rules based on second-order composition \( B^2 \):

$$
\text{(8) a. "Y saw X teach W to sing."}
$$

$$
\text{b. } ((S\setminus NP)\setminus NP) \setminus NP : \lambda w\lambda x\lambda y. \text{help}(\text{teach}(\text{sing}w)x)xy \equiv B^2 \text{sees}(B\text{teach sing})
$$

CCG thus reduces the operator \textit{move} of transformational theory to applications of purely adjacent operators—that is, to recursive combinatory \textit{merge}.

Interestingly, the latest “minimalist” form of the transformational theory has also proposed that move should be relabeled as an “internal” form of standard or “external” merge (Chomsky, 2001/2004: 110), though without providing any formal basis for the reduction other than identifying internal merge as “a grammatical transformation”. (If anything deserved the soubriquet “the lost combinator”, it would be this notional unitary combination of application and abstraction in a single perfect operator, linking or merging all types, as in the epigraph to this article.)

4.1 Expressivity of CCG

\( B^2 \) rules allow us to “grow” categories of arbitrarily high valency, such as \( (S\setminus NP)\setminus NP \). As we saw earlier, in some Germanic languages like Dutch, Swiss German, and West-Flemish, serial verbs are linearized using such rules to require crossing discontinuous dependencies. Thus, \( B^2 \) rules give CCG slightly greater than context-free power.
Nevertheless, CCG is still not as expressive as movement. In particular, we can only capture permutations that are what is called “separable”, where separability is related to the idea of obtaining the permutations by rebracketing and rotating sister nodes (Steedman, 2018).

For example, for the categories of the form $A|B$, $B|C$, and $D$, it is obvious by inspection that we cannot recognize the following permutations:

(9) i. *$B|C \ A|B \ C|D$
   ii. *$C|D \ A|B \ D|B$

This generalization appears likely to be true cross-linguistically for the components of this form for the NP “These five young boys”:

(10) i. *Five boys these young
   ii. *Young these boys five

21 of the 22 separable permutations of “These five young boys” are attested (Cinque, 2005; Nchare, 2012). The two forbidden orders are among the unattested 3.

The probability of this happening by chance is the prior probability of the hypothesis itself—that is the reciprocal of 24 choose 2—times 3 choose 2, the number of ways of predicting two impossible orderings out of three unattested ones, given by the following:

(11) $p = \frac{\binom{3}{2}}{\binom{24}{2}} = \frac{(3+2)/2}{(24+23)/2} = \frac{6}{552} \approx 0.01$

—that is, about one in a hundred. (If the sole predicted order that remains unattested so far were to be attested, this chance would fall to about one in two thousand.)

The number of separable permutations grows much more slowly in $n$ than $n!$, the number of all permutations. For example, for $n = 8$, around 80% of the permutations are non-separable.

There are obvious implications for the problem of alignment in machine translation and neural semantic parsing, to which we will return.

In 1986, I moved to Computer and Information Science at Penn, first as a visitor (incredibly, partly still under Sloan funding—at least, that was what Aravind Joshi told me), and then as a member of faculty, where much of the further development of CCG was worked out. Crucially, Aravind’s students proved in a series of papers (Vijay-Shanker, Weir, and Joshi, 1987, Joshi, Vijay-Shanker, and Weir (1991), passim) that the “shared stack” claim of Ades and Steedman...
(1982), was correct, by showing that both CCG and Aravind Joshi’s Tree Adjoining Grammar (TAG) were weakly equivalent to Linear Indexed Grammar (LIG, Gazdar, 1988), a new level of the Language Hierarchy characterized by the (Linear) Embedded Push-down Automaton (EPDA) (cf. Kuhlmann, Koller, and Satta, 2015).

The class of languages characterizable by these formalisms fell within the requirements of what Joshi (1988) called “mild context sensitivity” (MCS), which proposed as a criterion for what could count as a computationally “reasonable” theory of natural languages—informally speaking, that they are polynomially recognizable, exhibiting constant growth and some limit on crossing dependencies. However, the MCS class is much much larger than the CCG/TAG languages, including the multiple context free languages and even (under certain further assumptions) the languages of Chomskian minimalism, so it seems appropriate to distinguish TAG and CCG as “slightly non-context-free” (SNCF, with apologies to the french railroad company).

4.2 CCG for Natural Language Processing

Despite its SNCF complexity, CCG was originally assumed to be totally unpromising as a grammar formalism for parsing, because of the extra derivational ambiguity introduced by type raising and the combinatory rules, particularly composition. However, these supposedly “spurious” constituents also show up under coordination, and as intonational phrases. So any grammar with the same coverage as CCG will engender the same degree of nondeterminism in the parser (because it is there in the grammar). In fact, this is just another drop in the ocean of derivational ambiguity that faces all natural language processors, and can be handled by exactly the same statistical models as other varieties.

In particular, the head-word dependency models pioneered by Don Hindle and Mats Rooth, Mike Collins, and Eugene Charniak are straightforwardly applicable (Hockenmaier and Steedman, 2002b; Clark and Curran, 2004). CCG is also particularly well-adapted to parsing with “supertagger” front ends, which can be optimized using embeddings and LSTM (Lewis and Steedman, 2014; Lewis, Lee, and Zettlemoyer, 2016).

CCG is now quite widely used in applications, especially those that call for transparency between semantic and syntactic processing, and/or dislocation, such as machine translation (Birch and Osborne, 2011; Mehay and Brew, 2012), machine reading (Krishnamurthy and Mitchell, 2014), incremental parsing (Pareschi and Steedman, 1987; Niv, 1994; Xu, Clark, and Zhang, 2014; Ambati et al., 2015), human-robot interaction (Chai et al., 2014; Matuszek et al., 2013), and semantic parser induction (Zettlemoyer and Collins, 2005; Kwiatkowski et al., 2010; Abend et al., 2017). Some of my own work with the same techniques has returned to their application in
musical analysis (Granroth-Wilding and Clark, 2016; McLeod and Steedman, 2016), and shown that CCG grammars of the same SNCF class and parsing models of the same statistical kind are required there as well: it is only in the details of their compositional semantics that music and language differ very greatly.

Rather than reflecting on this substantial body of work in detail, I’d like to conclude by examining two further more speculative questions: The first is an evolutionary question: Why should natural language be a combinatory calculus in the first place? The second is a question about the future development of our subject: Will CCG and other grammar-based theories continue to be relevant to NLP in the age of Deep Learning and Recursive Neural Networks? I’ll take these questions in order in the next two sections.

5. Why is Language Combinatory?

Natural language looks like a distinctively Combinatory applicative system because \( B \), \( T \), and \( S \) evolved independently, in our animal ancestors, to support planning of action sequences before there was any language (Steedman, 2002). Even pure reactive planning animals like pigeons need application. Composition \( B \) and Substitution \( S \) are needed for seriation of actions. (Even rats need to compose sensory-motor actions to make plans.) Macaques can form the concept “food that you need to wash before eating”). Type-raising \( T \) is needed to map tools onto actions that they allow (affordances). (Chimpanzees and some other animals can plan with tools.) Second-order combinators like \( B^2 \) are needed to make plans with arbitrary numbers of entities, including non-present tools, and crucially including other agents whose cooperation is yet to be obtained. Only humans seem to be able to do the latter.

Thus, Chimpanzees solve the (mistitled) monkey and bananas problem, using tools like old crates to gain altitude in order to obtain objects that are otherwise out of reach (Köhler, 1925, figure 1a).

Köhler’s chimpanzee Grande’s planning amounts to composing \( B \) affordances \( T \) of crates etc., such as actions of moving, stacking, and climbing-on them, etc. Similarly, macaques can apply \( S(B \text{ before eat}) \text{wash} \) to a sweet potato (Kawai, 1965, figure 1b). Interestingly, Köhler and much subsequent work showed that the crates and other tools have to be there in the situation already for the ape to be able to plan with them: she was unable to achieve plans which involve fetching crates from the next room, even if she had recently seen them there.

More generally, the problem of planning can be viewed as the problem of search for a sequence of actions in a lattice or labyrinth of possible states. Crucially, such search has the same recursive character as parser search. For example, there are both chart-based dynamic-
programming and stack-based shift-reduce-style algorithms for the purpose. The latter seem a more plausible alternative in evolutionary terms, as affording a mechanism that is common to both semantic interpretation and parsing, rather than requiring the independent evolution of something like the CKY algorithm.\footnote{This point seems to have escaped Berwick and Chomsky 2016: 175-176,n9.}

Planning therefore provides the infrastructure for linguistic performance, as well as the operators for competence grammar, allowing the two to emerge together, as what was referred to above as an evolutionary “package deal”.

6. CCG in the Age of Deep Learning

I hope to have convinced you that CCG grammars are both linguistically and evolutionarily explanatory, as well as supporting practical semantic parsers that are fast enough to parse the web.

Nevertheless, like other supervised parsers, CCG parsers are limited by the weakness of parsing models based on only a million words of WSJ training data, no matter how cleverly we smooth them using synthetic data and embeddings. Moreover, the constructions they are specifically needed for—unbounded dependencies etc.—are as we have noted rare, and off in the long tail. As a consequence, supervised parsers are easy to equal on performance overall by using end-to-end models (Vinyals et al., 2015).

This outcome says more about the weakness of treebank grammars and parsers than about the strength of deep learning. Nevertheless, in the area of semantic parser induction for arbitrary knowledge graphs such as FreeBase (Reddy, Lapata, and Steedman, 2014), I would say that CCG
and other grammar-based parsers have already been superseded by semisupervised end-to-end training of Deep Neural Networks (DNN) (Dong and Lapata, 2016, 2018; Jia and Liang, 2016), raising the question of whether DNNs will replace structured models for practical applications in general.

DNNs are effective because we don’t have access to the universal semantic representations that allow the child to induce full CCG for natural languages and that we really ought to be using both in semantic parser induction, and in building knowledge graphs. Since the FreeBase Query language is quite unlike any kind of linguistic logical form, let alone the universal language of mind, it may well be more practically effective with small and idiosyncratic datasets to induce semantic parsers for them by end-to-end deep neural brute force (DNBF), rather than by CCG semantic parser induction.

But is it really possible that the problem of parsing could be completely solved by RNN/LSTM, perhaps augmented by attention/a stack (He et al., 2017; Kuncoro et al., 2018)? Do semantic-parsing-as-end-to-end-translation learners actually learn syntax, as has been claimed?

It seems likely that end-to-end SRL parsers and Neural Machine Translation (NMT) will continue to have difficulty with long range wh-dependencies, because the evidence for their detailed idiosyncrasies is so sparse.

For example, both English and French treat embedded subject extraction as a special case, either involving special bare complement verb categories (English) or a special complementizer “qui” (French).

(12) a. A woman who I believe (*that) won
   b. Une femme que je crois qui/*que à gagné

(This is actually predicted by CCG, which says that if you are an SVO language with relatively rigid-word order, like English and French, then you won’t in general be able to extract embedded subjects, whereas in verb initial languages like Welsh or verb final languages like Japanese and German, you will either be able to extract both embedded subjects and objects, or neither—Steedman, 1987, 2000.)

Not surprisingly, at the time of writing, a well-known end-to-end DNN translation system Near You shows no sign of having learned these syntactic niceties. Starting from a legal English, we get an ambiguous French sentence whose back-translation to English translation means something different:

(13) This is the company that the agency told us owned the title.
    ≠ C’est la compagnie que l’agence nous a dit détenir le titre.
    = This is the company that the agency told us to hold the title.
Similarly, if we start with French subject extraction, we get ungrammatical English (which happily translates back into the original correct French):

(14) C’est la compagnie que l’agence nous a dit qui détient le titre.
-≠ *This is the company that the agency told us *that holds the title.
= C’est la compagnie que l’agence nous a dit qui détient le titre.

If instead we start with legal English subject extraction, using a bare complement, the (correct) translation again includes an infinitival which is incorrectly back-translated:

(15) This is the company that they said had owned the bank.
-≠ C’est la compagnie qu’ils ont dit avoir possédé la banque.
= *This is the company they said they owned the bank.

By contrast, supervised CCG parsers do rather well on embedded subject extraction (Hockenmaier and Steedman, 2002a; Clark, Steedman, and Curran, 2004), which is a fairly frequent construction that happens to be rather unambiguously determined by the parsing model.

End-to-end methods are similarly unreliable when faced with long-range agreement, even when the source sentence is completely unambiguous in this respect:

(16) The banks think that the chairman owns the stock, and know that it is stolen.
-≠ Les banques pensent que le président est propriétaire du stock et sait qu’il est volé.
= Banks think that the president owns the stock and knows it is stolen.

Moreover, in principle at least, these constructions could be learned by grammar-based semantic parser induction from examples, using the methods of Kwiatkowski et al. (2010, 2011) and Abend et al. (2017).

These observations make it seem likely that there will be a continued need for structured representations in tasks like QA where long-range dependencies matter. Nevertheless, like Rock n’Roll, deep learning and distributional representations are clearly here to stay. The future in parsing for such tasks probably lies with hybrid systems using neural front ends for disambiguation, and grammars for assembling meaning representations.

7. The Shape of Things to Come

The most important open problem in NLP remains the fact that natural language understanding involves inference as well as semantics, and we have no idea of the meaning representation involved. If your question is Has Verizon bought Yahoo?, the text will almost certainly answer it many times over. But it is almost equally certain to present the information in a form that is not immediately compatible with the form of the question. For example, sentences like the following
use: a different verb; a noun rather than a verb; an implicative verb; an entailing quantification; a negated entailment; a modal verb; and disjunction:

(17) a. Verizon purchased Yahoo. ("Yes")
    b. Verizon’s purchase of Yahoo ("Yes")
    c. Verizon managed to buy Yahoo. ("Yes")
    d. Verizon acquired every company. ("Yes")
    e. Verizon doesn’t own Yahoo ("No")
    f. Yahoo may be sold to Verizon. ("Maybe")
    g. Verizon will buy Yahoo or Yazoo. ("Maybe not")

To arrive at a meaning representation language that is form-independent (and ultimately language-independent), we are using CCG parsers to machine-read the Web for relations between typed named entities, in order to detect consistent patterns of entailment between relations over named entities of the same types, using directional similarity over entity vectors representing relations. We then build an entailment graph (cleaning it up and closing it under relations such as transitivity—cf. Berant et al., 2015). Cliques of mutually-entailing relations in the entailment graph then constitute paraphrases that can be collapsed to a single relation identifier (Lewis and Steedman, 2013a). (This can be done across text from multiple languages—Lewis and Steedman, 2013b.)

We can then replace the original naive semantics for relation expressions with the relevant paraphrase cluster identifiers, and reparse the entire corpus using this now both form- and language-independent semantic representation, building an enormous knowledge graph, with the entities as nodes, and the paraphrase identifiers as relations.

To answer questions concerning the knowledge in this graph, we parse questions $Q$ into the same form-independent semantics representation, which is now the language of the knowledge graph itself. To answer the question, we use the knowledge graph and the entailment graph, and the following rule:

(18) a. if $Q$ or anything that entails $Q$ is in the knowledge graph then answer in the positive.
    b. If $\neg Q$ or the negation of anything that $Q$ entails is in the knowledge graph, then answer in the negative.

For this to work, we need complex expressions (including negation, auxiliaries, modals, implicative verbs etc.) to be nodes in their own right in both knowledge and entailment graphs. We shall then be able to discard hand-built knowledge graphs like Freebase in favor of a truly organic semantic net built in the language of mind, obviating the need to learn end-to-end transduction between semantic representations and the language of the knowledge graph.
If this project is successful, the language-independent paraphrase cluster identifiers will perform the function of a “hidden” version of the decompositional semantic features in semantic representations like those of Katz and Postal (1964); Jackendoff (1990); Moens and Steedman (1988); White (1994) and Pustejovsky (1998), while the entailment graph will form a similarly hidden version of the “meaning postulates” of Carnap (1952) and Fodor, Fodor, and Garrett (1975). Such semantic representations are essentially distributional, but with the advantage that they can be combined with traditional logical operators such as quantifiers and negation.

This proposal for the discovery of hidden semantic primitives underlying natural language semantics stands in contrast to another quite different contemporary approach to distributional semantics that seeks to use dimensionally-reduced vector-based representations of collocations to represent word meanings, using linear-algebraic operations such as vector and tensor addition and multiplication in place of traditional compositional semantics. It is an interesting open question whether vector-based distributional word embeddings can be used, together with directional similarity measures, to build entailment graphs of a similar kind (Henderson and Popa, 2016; Chang et al., 2018). It is quite likely that some kind of hybrid approach will be needed here too, to combat the eternal silence of the infinite spaces of the long tail.

8. Conclusion

Algorithms like LSTM and RNN may work in practice. But do they work in theory? In particular, can they learn all the syntactic stuff in the long tail, like non-constituent coordination, subject extractions, and crossing dependency, in a way that will support semantic interpretation? If they aren’t actually learning syntax, but are instead learning a huge FST or a Soft ATN, then by concentrating on them as mechanisms for natural language processing, we are in danger of losing sight of the computational linguistic project of also providing computational explanations of language and mind.

Even if we convince ourselves that something like SEQ2TREE is a psychologically real learning mechanism, and that children learn their first language by end-to-end mapping of the sentences of their language onto the situationally-afforded structures of the universal language of mind, we still face the supreme challenge of finding out what that universal semantic target language looks like. We won’t get an answer to that question unless we can rise above using SQL, SPARQL and hand-built ontologies and logics such as OWL as proxies for the hidden language of mind, to use machine learning for what it is really good at, namely finding out such hidden variables and their values, for use in a truly natural language processing.
Acknowledgements

The work was supported in part by: ERC Advanced Fellowship GA 742137 SEMANTAX; ARC Discovery grant DP160102156; a Google Faculty Award and a Bloomberg L.P. Gift Award; a University of Edinburgh and Huawei Technologies award; my Co-Investigators Nate Chambers, and Mark Johnson; the combinator found, Bonnie Webber; and all my teachers and students over many years.

References


Kuncoro, Adhiguna, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In Proceedings of the 56th Annual Meeting of the Association for


Computational Linguistics, Volume 44, Number 4


S. Szabolcsi. 1989. Bound variables in syntax: Are there any? In Renate Bartsch,


