BOOK REVIEW

IS COMPUTATIONAL A DEAD LINGUISTICS

REVIEW OF "COLLECTED PAPERS OF MARTIN KAY: A HALF-CENTURY OF COMPUTATIONAL LINGUISTICS"

Seán O Nualláin

Martin Kay with the editorial assistance of Dan Flickinger & Stephan Oepen, Collected Papers of Martin Kay: A Half-Century of Computational Linguistics, CSLI Studies in Computational Linguistics. Center for the Study of Language and Inf, 2010, 639pp. ISBN: 9781575865713 ~ Hardcover. £31.50, $35.00.

Computational linguistics (CL) should be one of the world's hottest academic subjects. It attempts to add a computational characterization of language to the formal characterization achieved in linguistics proper. It adds components of computer science, cognitive psychology and logic to students hungry for an interdisciplinary degree. Moreover, it is the natural locus for applied work in software localization, a process of adapting software not just to a language but a whole culture, adding further excitement to the student experience.

As if that wasn't enough, there is a chronic worldwide shortage of professionals in the area of software localization, making it a sure-fire opportunity for a career. In 2012 Louise Phelan, PayPal's vice-president of operations for Europe, the Middle-East and Africa, and winner in 2014 of the title of Ireland's most trusted leader, claimed that she had to import 500 software localization professionals for Ireland alone for Paypal alone. Informally, similar figures have been adduced by SAP and others.

Kay's book affords a chance to see what went wrong. The fact that, a la Wittgenstein, he submitted this oeuvre for his Ph.D. rather than actually writing a thesis may tell the reader something; British computational linguistics from the 1960's was a bloodbath of failed Ph.D's, with the unsuccessful applicants ending up running publishing companies, writing texts, and in one memorable case turning things around only after it became clear to the dean of Graduate studies, reading through the assessment, that the examiner was insane.

Thus, this review will begin with the institutional problems with the discipline. It will proceed to parse the intellectual difficulties that Kay exemplifies. Briefly, Kay – an admirer of Halliday - stresses that language encodes and transmits ideas. However, he shows little feel for cognition, and the stronger chapters in this immense book are centered on his computational contributions.

Kay writes well, and often acerbically; gargantuan projects like the disastrous EUROTRA attract his ire (552);

> "The EUROTRA project began….as a massive attempt to build a practical fully automatic translation system"

That was not to be;

> "It ended lamely attempting to justify the expenditure of close to a hundred million ECU on the grounds that researchers even in the poorest European nations had learned something of Computational linguistics" (ibid.)

Kay leaves out what happened next, which should be a cautionary tale for all current massive investment in Big Science. What became of these researchers? They had enjoyed up to a decade of jet-setting; in fact, they nicknamed their project "EUROTRAVEL". They had sloppy or non-existent work habits after spending what should have been their peak working years essentially learning to be courtiers in a project that never delivered. *Pace* the EU, there HAD been research going on already even in those very poorest European nations. In fact, the arrival of EUROTRA alumni often destroyed degree programmes as well as collegial relationships.

When recessions hit, as they do with increasing frequency, universities are always asked to trim fat. One method is full-frontal assault on tenure; that always ends badly, after management is allowed to fire a few token miscreants (as management see them). A much safer target is interdisciplinary degrees like computational linguistics that are typically shared between departments like computer science and linguistics, neither of whom is heavily invested in it. Consequently, the mini-recession after 2001 and the rather more serious one since then resulted in a cull of programmes.

Intellectually, Kay inveighs against the idea that "translation was treated as a purely linguistic problem" (ibid) in systems like EUROTRA and the later Verbmobil.

This he regards as a mistake; better, he argues, to get a domain expert fully to characterize what the content of a document is. While the overhead for translation between 1 source and 1 target language might be high, there is an economy of scale that will pay off, particularly in today's world with companies like SAP dealing in over 100 languages.

Paradoxically, Kay continues (523), the more linguistics, the worse the automatic translation (MT) system. SYSTRAN, relatively innocent of linguistics and a result of a US project from the 1960's, is still in many ways the gold standard, as he argues (ibid). That has left the door open for Google to do away with linguistics altogether, and crawl the web incessantly looking for statistical matches. While Kay in many ways anticipated this work, he clearly does not approve of it.

What does he approve of? It is as well to say that this book is a superb production; hardly a proofing mistake in sight, fine binding and -as we shall see- historically interesting content. We shall start with his elaborated views on MT. We shall go on to his seminal work on chart parsing, unification and other computationally-tractable ideas. His vision, on the one hand, is general enough to subsume all of language and computationally specific enough, on the other, to recommend that a LISP "CONS" cell is an appropriate implementation for a specific idea.

We shall end with the rock on which the whole ship founders; context, and Kay's echoing of a whole generation in his inability to deal with it. In particular, to anticipate, his unwillingness to handle sublanguage is glaring.

In a rather unfortunate analogy, Kay likens the process of MT to the legs of one's trousers. One is on the source, the other on the target language. Around the crotch (not usually the center of cogitation), linguistics gives way to psychology, as mental representations become causal and important. However, one can transfer around the knee by linking the source syntactic representation to the target syntactic representation; or around the ankles, a la Google, the late Jelinek of IBM and others, by simple statistical correlation of words.

If the latter technique worked, we could all go home; the fact is that it doesn't and never will. There are numerous reasons why it doesn't, including the fact that humans would get very bored with a language that functioned like that and would use something more interesting. In fact, attempts to impose languages like Esperanto without grammatical irregularities have shown that humans need something meatier.

Google's limited successes have used techniques like "deep learning" that – sometimes brilliantly – try to emulate neural processing by treating cognition as a parallel process with "shallow" computation (that is, a very few steps, in line with the relatively slow pace of neural transmission). Yet these successes are in general less

impressive than the Meteo system of the University of Montréal, which went live in the mid-70's correctly translating weather forecasts between English and French.

Kay argues that this domain of weather forecasts is a sublanguage, and MT can work for such; but we lack any rules for what constitutes sublanguage. He is rather too pessimistic; while he repeatedly refers to Kittredge's work, he seems unaware that Jerry Hobbs defines sublanguages as a degree of restriction of context in which selectional restrictions, usually the province of semantics, become that of syntax. So, while we normally use semantics to ascertain that only a human can form an intent, in a sublanguage that restriction would be handled by the syntax. Moreover, Wittgenstein famously argued that there was a further restriction of context in which the single word "Slab!" is a synonym for "Hand me a slab".

Kay's major contribution is in limpid formalisms to handle syntactic parsing, particularly in fixed word order languages like English. Chart parsing, like Mitch Marcus' lookahead parsing, is elegant and tractable, particularly in LISP. In this book, Kay is also exercised with description at the level of ideas, and gives much space to the corresponding "functional" grammars. Both formalisms are simultaneously handled with an operation called unification, an extension of set union. "Functional" grammars can be extended to (apparently) free word order languages like Finnish, a point Kay explores with Karttunen,

This writer is unconvinced that these systems can scale. He is also unconvinced of the sincerity of Kay's stated adherence to Halliday's view of language. The strongest sections in this book are the computational descriptions; the weakest are the gestures toward a cognitive theory of language. It may be the case, as I explore elsewhere (2003, 2014) that we are missing something; while the term biolinguistics has been taken to mean an evolutionary account of language, within the biosemiotics community there is a move afoot toward a truly general account of symbolic functioning, one that encompasses all its manifestations in biology.

Finally, techniques like word2vec (Manning et al, 2014) successfully fulfill Kay's program of vector representation. Useful commercial systems like Stitchfix are already in place. Ironically, sublanguages admit of many different solutions. The fact that "deep learning" – to repeat, a small number of computations, in parallel – has also been employed for semantic representations gives pause. Finally, speech processing has been so thoroughly solved at Google that there now are security problems; your smartphone can be commandeered by a message over an intercom intelligible only to the machines we carry.

So therefore is computational linguistics dead? It seems true that sentiment analysis and other expressions of mass opinion are better handled by word2vec and its

peer technologies. Problems arise when trying to model a specific user's meaning, be that a novel, a software manual, or a poem. Computational linguistics can make a strong case for itself, one only partly based on the elegant formalisms in this book which may not be quite the breakthroughs we one thought they were.

president@universityofireland.com

REFERENCES

Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

O Nualláin Seán (2003) "The Search for Mind" (Third edition England: Intellect)

O Nualláin Seán (2014) "Symbolic and Cognitive Theory in Biology" In Cosmos and History: The Journal of Natural and Social Philosophy, Vol 10, No 1 (2014)