Computational linguistics beyond the processing of English

Diana Santos[*]

27 August 2007

Abstract

Processing of the English language is overwhelmingly mainstream in computational linguistics. This text claims that this situation is neither healthy for computational linguistics nor theoretically tenable.

1 Introduction

One of the most interesting (empirical) facts of linguistics is that there are many natural languages. Computational linguistics – either taken as the use of computational tools to do linguistics, or as performing language tasks with a computer – does not deal properly with this situation. In fact, it hardly deals with it at all, because the overwhelming bulk of work is done on the processing of English.

Imagine that only one disease were studied and published on medical jounals, or one kind of drug in pharmaceutics, because doctors firmly believed that all diseases were the same, or pharmacists that all drugs had the same effect. Surely one could not expect that significant progress would take place in either science? In fact, doctors cannot proceed on the basis "if this is so for rabies, then it must be so for cholera!", and neither can pharmacists hold that "if this pill works for typhoid, then it works for meningitis as well" without a strong empirical basis for their hypotheses. However, it appears

[*] Linguateca, SINTEF ICT, Pb 124 Blindern, N-031 Oslo, Norway. E-mail: Diana.Santos@sintef.no

to be common place in computational linguistics to utter “if this is true of/works for English, it must be true of any language as well” and proceed accordingly.

In my view, this is methodologically outrageous, completely undermining the view of (computational) linguistics as an empirical science or proper engineering.

This text will claim that languages are different on two accounts, both of them relevant to our discipline. I start by pointing out some basic undeniable facts and try to show that they lead to the conclusion that each language has to be dealt with separately; then, I make the bolder claim that to prove or disprove my assertion that languages are genuinely different in what they convey – which not everyone would agree with – the only methodologically sound way to proceed is to make no assumptions, and test whether, in the end, languages serve (or do not) the very same ends.

2 A gentle overview of language differences from a very practical point of view

For the moment, let us put aside any philosophical objections, and simply underscore the following rather commonsensical observations:

- For language learning, proximity of languages is often a factor of significance. Closely related languages are easier for humans to learn than totally unrelated ones. Also, to focus on the specific differences between the foreign language and the learner’s native one is a common pedagogical tool.

- Frequency issues are commonly considered an important property of language, of consequence both for language development (e.g. grammaticalization) and for language understanding (in infants). Now, it is well known that frequency lists (of words, lemmata or grammatical constructions) differ from one language to another.

- Grammar (for example, what is obligatory and what is optional) also varies widely among languages. In fact, in this respect it is even more enlightening to see how the grammar of even closely related languages differs. As soon as something is obligatory, speakers have to make choices, even if they might not wish to. Their language forces (automatic) attention to features or concepts to which other languages simply do not pay attention.

- The importance of metaphor in language is now almost universally acknowledged. It is known that different languages function with different metaphors, therefore structuring knowledge and communication in different ways. One clear example is spatial reasoning, which, although basic in most natural languages, differs widely in the way it is performed. A possibly degenerate example of this is the crucially different use of spatial prepositions (or cases) when it comes down to non-obvious cases.

All these observations seem to be good indications that languages are different, and common sense would have it that CL researchers should devote a lot of work to deploy systems that work for different languages, and deal properly with each language in its own terms.

However, sadly this is not so. What most people do when they are not working on English is to take (developed for English) things off-the-shelf and adapt them to their own language, without even questioning the rationale of doing so. Then, they evaluate those tools in tasks devised for English, again without questioning whether the tasks themselves make sense or are well defined for their language. As I claimed in [5], when people apply techniques and tools developed for English to other languages, they succeed or fail exactly to the extent the langage is similar or different from English. This is not very enlightening, if one expects or believes one is processing X, and not English. As a result, we enter a vicious circle of English based language processing from which it is impossible to extricate ourselves.

Let me adduce some examples from both computational linguistics and linguistics proper (if this distinction makes any sense):

- **Lexical ontologies.** Last year I had a student presenting his project of creating a lexical ontology for Portuguese to a Norwegian audience [7]. A heart-felt question or comment was: “Why Portuguese? Can’t your work be more general?” I bet that George Miller and his colleagues were never confronted with the question/objection in WordNet’s [4] early days (or ever): “Why English?”

  On the contrary, I even suspect that many NLP practitioners – especially native English speakers, but sadly not only – expected the opposite: now that WordNet exists, it exists for all languages (merely a translation nuisance), and that it has come as a (bad) surprise that so many people seem to be "redoing" the work or starting from scratch.

In fact, many authors still excuse themselves (or boast) that they are using English WordNet for their own language, due to the unavailability of the right one. This shows that they missed the whole point of WordNet as a description of English (and not natural language) categorization: for any person acquainted with WordNet knows that it is the empirical facts of English, not "natural categories" of any sort that define WordNet's synsets.

- **Event classification.** Another striking example of misappropriation of excellent research for English into a general semantic framework is the Vendlerian classification of events (achievements, activities, accomplishments) which [10] founded and explained in terms of English grammatical categories. It has been imposed without crticism to all other languages, as extensively discussed in [6]. While Vendler explicitly states that he is dealing with English, and even discusses the philosophical question of using language data to do philosophy in [11, page 5, my emphasis]:

  > even the linguistic data giving the structure of **a particular natural language** are a fruitful source of genuine philosophical insight

  we have seen this framework being applied (or simply adapted) to almost every other language on the planet!

- **Question answering.** This year at QA@CLEF, a crosslingual non-English monolingual question answering evaluation contest, participating systems had to answer questions grouped by topic, and formulated in a natural way, which meant anaphoric reference and context-based questions. Even though the questions themselves were purposely fairly similar, there was a wide range of differences when the four languages German, Spanish, French and Portuguese formulated the questions [1]. Clearly, developers of QA for different languages will have to tackle different problems, and adapt their systems to the way people ask questions in their own languages, even if they are all addressing the "same" problem.

In fact, few people would deny that different languages pose different problems to their computational processing (even if there were an underlying deep similarity). But they tend to forget that this logically requires different attention and effort in different areas for different languages in CL, as [9, page 144] say so well on the issue of creating evaluation resources:

> data has to be collected for different languages, and the data has to be comparable: however, if data is functionally comparable it is not necessarily descriptively comparable (or vice versa), since languages are intrinsically different

In other words, the task of identifying named entities would never be singled out for processing German, nor would part of speech tagging be proposed as a sensible module for Portuguese, although they made sense for English. Tokenization or lemmatization are hard for Chinese and Hebrew, but are fairly unproblematic for English. Classifiers, aspect and compounding, although it is possible to find some (small) counterpart of those phenomena in English, according to some scholars at least, are obviously not of utmost relevance as compared to Japanese, Russian or Norwegian respectively. I think it is fairly evident that no work on those matters in English alone will be able to compensate for the effort required for these languages.

Let us consider the matter the other way around. If anyone does good work on English processing, it is natural that they try to generalize their technique to "language" – but this is often done without using proper criteria: people consider the lexicon, or grammar, or both, as black boxes, and then generalize the system to all languages, without due investigation into whether the task still makes sense, the users still need it, and so on. In fact, this "abstraction from English" procedure will make sense only to the extent the language is similar to English.

Then there is yet another possible twist: one can construe "language" by transferring most of English into it but disguising it as an abstract concept (as [8] points out, this is often the case in knowledge representation formalisms which use "English in disguise"). Again, any other natural language will only be translatable into that formalism when it is similar enough to English, and we have not advanced the computational processing of that language except for (trivial) machine translation into English.

3 A deeper problem

But now, imagine that it is not only a question of different details requiring more processing than others, but that languages are also different in the meaning they convey. For an excellent defence of such a position I refer the reader to [2], and here I shall simply invoke [3, page 166]'s remark as to how improbable it is that languages should convey the same thing:

> it would surely be surprising, and a very strong empirical claim, that different languages using different means to express 'meanings' always arrived at exactly the same end.

In fact, there has never been any attempt to empirically confirm such a belief, and I think it is high time to stress that we have strong empirical evidence to the contrary. In translation, even when, in principle, one might be able to say "the same", human translators tend to make choices that either add or remove content in order to conform to what is expected and implicit, or must be made explicit in different languages. So far, there is no (computational) system I know of that attributes the same meaning to two sides of a human translation (or even attempts to do so).

All this should easily cast enough doubt on the possibility of maintaining such a belief without any attempt at providing evidence for it. If people who do computational linguistics adhere to such a creed, it is high time that they should take this challenge seriously.

I am well aware that people who are convinced that "languages are just a different arbitrary code for the same message", which logically entails that one would then simply transfer whatever is done for English to the language of one's choice (or vice-versa), will never take the above examples – or any similar in kind – as relevant, because they work in a different paradigm in Kuhn's sense. For them, what is interesting is the abstract similarities, what is common to all languages despite the differences (themselves a nuisance). Specific details are swept aside as uninteresting, even if they happen to concern English. And thus, once done (for English), there is nothing else to publish on.

My (precisely opposite) position – namely, that features that are interesting, or relevant to language processing, are not in any way related to features that are "common to all languages", and that we should accept that different languages say different things, and so process each language independently – has two specific advantages:

1. these differences can be empirically validated or refuted – in other words, they are falsifiable;

2. and the process ensures, from the start, a perfect fit for each language.

4 The current situation is too biased

Due to many unrelated factors, and also to a kind of snowball effect (the more research is done on English, the more interesting and publishable is the work done on English), English is uncontroversially the language of computational linguistics. This is too narrow for a discipline that deals with handling natural languages with computers, and leads to a severe bias in what is investigated and implemented that would not occur if languages were treated (as scientific objects of study) in an equal way.

We have seen that this is partially due to the common assumption that, at a deep enough level, the same is, or can be, conveyed in any language. And consequently, it is enough to do CL for one language, while others are trying to uncover the way to go automatically to the other languages. However, let us pause to admit that nobody has yet found that way.

One often hears a totally different kind of claim: that adapting work for English to other languages is more cost effective than developing things from scratch. But I have never seen such claims being backed by empirical data. On the contrary, they are simply used as an argument to go that way in the first place.

I believe it is time that we, as practitioners of this discipline, looked at other languages without English eyes. That will be beneficial – I believe – even to English language processing, because it is possible that different problems or solutions will turn out to be applicable to English as well.

It is already far from irrelevant that one needs, in order to achieve international visibility in CL, to publish **in English** in our discipline (with the automatically implicit preference for native speakers as authors). It is far too much to require, and harmful to the area as a whole, that one must publish **on English** as well.

Acknowledgment

This work was done in the scope of the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC.

I am grateful to Nuno Cardoso, Marcirio Chaves, Luís Costa, Jan Engh, Ana Frankenberg-Garcia, Belinda Maia, Cristina Mota, Nuno Seco and Mário J. Silva for pertinent comments on a previous version, and to Belinda again for considerable help in improving the English.

References

[1] Luís Miguel Cabral, Luís Fernando Costa, and Diana Santos. Esfinge at CLEF 2007: First steps in a multiple question and multiple answer approach. In Alessandro Nardi and Carol Peters, editors, *CLEF 2007 Working Notes*, 2007.

[2] John M. Ellis. *Language, Thought and Logic.* Northwestern University Press, Evanston IL, 1993.

[3] Edward L. Keenan. Some logical problems in translation. In F. Guenthner and M. Guenthner-Reutter, editors, *Meaning and Translation: Philosophical and Linguistic Approaches*, pages 157–189. Duckworth, Manchester, 1978.

[4] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[5] Diana Santos. Towards language-specific applications. *Machine Translation*, 14(2):83–112, June 1999.

[6] Diana Santos. *Translation-based corpus studies: Contrasting English and Portuguese tense and aspect systems.* Rodopi, Amesterdam/New York, 2004.

[7] Nuno Seco. Building a Large Scale Lexical Ontology for Portuguese, SINTEF, 16 August 2006 2006.

[8] Sergei Nirenburg and Yorick Wilks. What's in a symbol: Ontology, representation and language. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(1):9–23, 2001.

[9] Karen Sparck-Jones and Julia R. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review.* Springer, 1996.

[10] Zeno Vendler. Verbs and times. *The Philosophical Review*, LXVI(2):143–160, 1957.

[11] Zeno Vendler. *Linguistics in Philosophy.* Cornell University Press, Ithaca, 1967.