DISTORTION OR IMPROVEMENT
– Effects of information technology on the development of natural languages

Hans Karlgren
Kval
Skeppsbron 26
S-111 30 Stockholm
Sweden
+46 8 715 32 46 (voice)
+46 8 715 36 75 (fax)
karlgren@sics.se

If we do not yet have an answer, let us hope that the question is premature.

On some occasions during Coling conferences we should raise our gaze from the hows of our trade to regard the whys and whynots.

Ever since the first Coling was arranged and the term computational linguistics was coined, our undertaking has had two faces: research, where computation is used for better understanding of language, and application, where understanding of language is used for better computation. It has been taken for granted that practical effects, if any, are positive. In particular, it has been tacitly assumed, the languages in use between humans, what engineers see as some unique next to metaphysical entity which they call "natural language" - in singular! - would be unaffected. Is that really so?

Of course, human/machine interaction creates new genres: we adapt to whoever or whatever is our interlocutor, that is a well-known feature of our linguistic competence. Whether or not they claim to be "NLP", machines typically behave conspicuously differently than humans in communication, employing either some explicitly declared machinese or some unspecified superset of some unspecified subset of some human language. This fact will presumably remain true in the next century. Humans adapt accordingly. The machinese-induced human habits will spread to human/human situations, and more so as machines get better and more centrally placed in the community. - Besides, the distinction between man/machine and man/man tends to be blurred as communication becomes machine-mediated to an increasing extent. Will there remain any significant amount of writing, say a century from now, which is not at some step machine-supported?

Which are the long-time effects of exposure to machinese or machinese-like language? More consistent and less ambiguous use of terms and phrases? Greater clarity of sentence, discourse and argument structure? And/or: Decay of finer shades of meaning, emotional overtones and the social subtleties between information and command, between "I tell you" and "I tell you to"? Damping of innovation, humour and irony? Esthetic deterioration?

Empirical evidence is scanty, of course. We are only at the beginning of a beginning. Can we note more than mild tendencies towards stereotypisation and, slightly more ominously, an increasing verbosity?

In some fields we can sense the writers' anxiety to catch the eye of current poor retrieval systems. Established formulations are preferred, particularly in titles, to please not the reader but the computerised reader's digest. Thus, in

some branches of legal informatics, the efforts made by today's writers to attract tomorrow's readers by paving the way for yesterday's retrieval technology has a stultifying effect.

We can also note how writers intentionally simplify their style to make it amenable to existing unsophisticated translation tools. That has an impoverishing effect - not necessarily worse, though, than the self-imposed discipline of writers mindful of human translators or non-native readers who are less familiar with the language used.

Is the distortion of human language a phenomenon of a passing phase? Will improved language technology reduce the effects by narrowing the gap between human and machine competence? Or rather promote mechanisation, making programmed agents more influential? Are certain kinds of technological advances particularly urgent because of their consequences for human usage? Or, conversely, are there particular features of artificial systems which we should refrain from introducing because of such side-effects even though they may be cost-effective for their immediate purposes?

The impact on normal human behaviour may be intended.

Some countries - the Scandinavian countries being very clear examples - have a national language policy with rules and recommendations generally taught and widely accepted, to promote consistency, clarity and continuity combined with adequate doses of motivated changes. Many private companies, publishing houses, newspapers and other organisations have an elaborate house policy on style of writing and speaking. Typically, such regulation refers to low linguistic levels: spelling, terminology, name forms, certain phrases and headings, document layout &c.

One reason for the preoccupation with editorial detail is the legislators' lack of linguistic sophistication: they do not possess the intellectual tools to describe the desired text properties, nor have linguists, if consulted, so far had much of substance to offer on higher levels of structure.

The other reason is technical. Rules on higher level than spelling and vocabulary have been hard to enforce. Large-scale supervision has not been feasible. We are just beginning to see non-trivial computational support alerting writers and editors of, e.g., provincialisms (such as Americanisms in British English and vice versa), "Eurospeak" jargon in the European national languages, *he* when women are meant to be included, high-brow words and style when a broad audience is addressed.

Now, as computational linguistics will provide the intellectual and operational tools it will be possible to specify and enforce style guides worth the name.

If virtually all text is crunched through the same corporate, national or global network mill, norm adherence and standardisation can be warranted on a many levels. Exhaustive relevant in-depth real-time analyses will be practically and economically justifiable on as large text flows as the population could possibly type or pronounce. Deviants can be automatically identified, commented upon, amended, returned, censored and/or punished (say, in excusable cases, by some intentional delay).

That sounds like a brave new world for those who care about language development. When writer's support and verification tools have gone far beyond present myopic spellers and grammar checkers, text production may become a controlled process, where the needs for clarity, consistency, continuity and innovation are skillfully balanced by uniquely well-informed planners with effective tools at their command to steer development along rational lines. Will computational linguistics contribute to

enhancing human communication above anything we have so far imagined?

And/or shall we expect a trend towards constrained languages in a more controlled society?

Will computational linguistics play a key role in changing human languages and the rules of the language games in society? If not in a decade or two, then in a remote but conceivable future?

As toolmakers, is it our duty to suggest opportunities and issue early warnings as well as to supply on demand?