The New Paradigm in NLP and its Impact on Translation Automation

Elliott Macklovitch
RALI – Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, Canada

1. Introduction

Allow me to begin on a personal note. I have been actively involved in machine translation (MT) for nearly 25 years – since 1977, to be exact. That is a long time by anyone's standard, particularly in a field that is barely 50 years old. A great deal has changed over the period of my involvement in MT. In this paper, I would like to step back and try to analyse some of those changes with a view to identifying the major trends, especially those that have emerged in recent years. Hopefully, this exercise will allow us to discern a little of what the future holds in store, both for professional translators and for the translation industry.

The theme of this Symposium is the impact that new information technologies and the Internet are having on the work of terminologists, translators, technical writers and others in the language community. Now to be perfectly frank, I'm not altogether certain what is encompassed by the term "new information technologies"; what is more, I won't have a great deal to say about the Internet. Rather, I propose to focus on another equally important, although perhaps more technical issue: the fundamental paradigm shift that has occurred over the last ten or fifteen years in the general field of natural language processing (NLP), of which MT is one sub-domain, and the far-reaching effects that this has had on efforts to automate the process of translation. I also want to examine what may reasonably be viewed as progress in MT over the same general period, both in the methods employed to develop MT systems and in the quality of the results produced by these systems. For contrary to what may be the general belief, I am of the opinion that MT has made considerable progress since I first innocently wandered into its confines some 25 years ago.

2. Machine translation then

Let us begin with the overall objective of the enterprise as it was construed back then. For the major part of its fifty year history, MT's basic objective has been the automation of the entire translation process. Ideally, MT meant taking a text in one language, running it through a computer system and automatically obtaining an equivalent text in another language, perfectly grammatical and idiomatic, with all the meaning of the original intact. Computer-assisted translation (CAT) was an unknown concept during MT's early years; and in a sense, CAT only emerged with the recognition that fully automatic, high quality translation – or FAHQT, as (Bar-

Hillel 1951) first called it – was by and large unattainable for unrestricted texts, and would remain so for many years to come.

Today, this optimistic view of MT's potential to wholly supplant the human translator may strike us as rather naive. In large part, I think it can be explained by the fact that most of the people working on MT development in those early years were engineers and computer scientists, i.e. people with no direct experience of translation and little understanding of what an enormously sophisticated and complex process it is. One notable exception to this generalisation was Martin Kay, a pre-eminent researcher not only in MT but in computational linguistics in general. For years, Kay doggedly warned people of the folly of attempting to mechanise a process, essential elements of which remained largely mysterious.[1] In 1980, he published a remarkable article entitled "The Proper Place of Men and Machines in Language Translation", which was recently republished in (Kay 1997). In the following memorable passage, Kay caricatures – though only slightly – the manner in which MT was then being employed by the U.S. military and intelligence communities:

> "There was a long period – for all I know, it is not yet over – in which the following comedy was acted out nightly in the bowels of an American government office with the aim of rendering foreign texts into English. Passages of innocent prose on which it was desired to effect this delicate and complex operation were subjected to a process of vivisection at the hands of an uncomprehending electronic monster that transformed them into stammering streams of verbal wreckage. These were then placed into only slightly more gentle hands for repair. But the damage had been done. Simple tools that would have done so much to make the repair work easier and more effective were not to be had… In short, one step was singled out of a fairly long and complex process at which to perpetrate automation. The step chosen was by far the least well understood and quite obviously the least apt for this kind of treatment." (Kay 1997, p.5)

2.1 From the users' perspective

Although Kay doesn't explicitly name them here, those designated to do the repair work on the raw machine output in this, the classic model of MT collaboration, were almost always human translators, and the repair work itself was generally known as post-editing. This is a second major characteristic of MT as it was practised 25 years ago: with the notable exception of the intelligence community, which could sometimes make do with unrevised output for information scanning purposes, the targeted users were almost invariably professional translators. For years, MT system vendors attempted to convince translators and translation service managers to implement the basic arrangement described above, in which the machine would first produce a rough draft that translators would then post-edit until they deemed it to be of deliverable quality. For years, MT vendors maintained that this arrangement could be made to work cost-effectively – if only the translators would overcome their aversion to (or fear of) computers and just continue updating the system's dictionaries. Yet, despite all the pressure to "modernise", most translators remained sceptical. And in fact, the more rigorous MT evaluations

[1] Translators often invoke the notion of "context" to justify the selection of one target equivalent over another, although they cannot always explain precisely which contextual elements are crucial in any given case. Kay's point was that unless this and other such notions can be *formally* defined, there is no way a machine can hope to automate them.

demonstrated that they were correct to be reticent. For implementing MT in this way to produce high-quality translation of unrestricted texts rarely turned out to be cost-effective.[2]

The MT vendors eventually revised their strategy and began to present their systems as technology that was complementary to translators' skills. These systems, they argued, would take care of the donkey work, removing the tedium from translation and allowing the translator to concentrate on the really challenging problems. Indeed, if that were the case, it is hard to understand why translators would resist. And if they did resist, often with tenacity, it was because they had concluded that post-editing MT output did not accelerate their production or improve their quality of life. On the contrary, having to wrestle with "stammering streams of verbal wreckage" was perceived as a degrading task, in the literal sense of the word: it brought the human down to the level of a dumb, i.e. uncomprehending machine. For that was and still remains MT's fundamental challenge: comprehension. Producing an accurate and coherent translation routinely requires attaining a "deep" understanding of the source text, which in turn requires an extensive and largely unpredictable range of both linguistic and extra-linguistic knowledge, as well as the ability to reason over that knowledge. Human translators achieve this kind of understanding almost unreflectingly, because they are intelligent in ways that machines are not; among other things, humans bring to this task an innate knowledge of language, not to mention a lifetime of experience and culture. Machines may be able to calculate at lightning speed, but they have no culture and only as much knowledge as humans have been able to feed them; hence their understanding of natural language texts is necessarily quite limited. As any qualified translator will tell you, attempting to translate a text – even a dry, technical text – with less than a full understanding can be dangerous and often disastrous. If this is true for humans, how much more so for machines!

Of course, the situation is not quite as simple as this; for the usefulness of a machine translation is not an all or nothing question, but rather a matter of degree. Moreover, it *is* at times possible to produce an acceptable translation on the basis of a limited understanding of the source text, particularly between closely related languages, like English and French. Nevertheless, time seems to have proven those reluctant translators at least partially right. For one of the more significant changes in the MT landscape is that almost all the vendors of general-purpose MT systems who targeted translation services – whether it be Weidner, ALPS, METAL or Tovna – have since gone out of business.[3] Those companies, like Systran, that still remain are doing a different kind of MT business nowadays. And those few systems which are still being used with success by internal translation services, like the SpanAm system at the Pan American Health Organization, have been gradually developed and groomed by the same people who exploit them. Significantly, no translator at PAHO or at the Translation Service of the European Commission is forced to work with MT output; those who freely choose to do so are convinced that it is worth their while.

2.2 From the developers' perspective

Between 1977 and 1981, I had the privilege of working as a linguist at TAUM, the University of Montreal's machine translation project. Founded in 1965, TAUM was long

[2] For the report of one such trial, see (Macklovitch 1986).
[3] With the exception of those in Japan, which is a very different market.

considered to be one of the premier centres for MT research in the world. It was at TAUM that the prototype of the Météo system was first developed. For those who don't know it, Météo is the specialised MT system which translates all of Canada's weather bulletins between English and French, in both directions: over 45 thousand words a day, requiring minimal post-editing.[4] This system is justly considered by many people to be one of MT's undisputed success stories. Following the delivery of the Météo prototype, the federal Secretary of State commissioned the TAUM group to develop another specialised system, this one for the maintenance manuals of a coastal patrol aircraft which the federal government had just purchased; a task far more difficult than weather bulletins.

The TAUM-Aviation system, as it came to be called, has been described in detail elsewhere, c.f. (Isabelle 1987). The system itself was never actually put into operational production, for reasons that are too complicated to go into here. Nevertheless, TAUM-Aviation was without a doubt one of the most sophisticated and ambitious MT systems in existence at the time, and the TAUM group an undisputed world leader in MT research and development. I would like to briefly describe what is was like to work on an MT project in those years, in order to allow us to better appreciate to what extent things have changed today.

At the height of the TAUM-Aviation project in the late 1970's, there were about twenty-five highly trained specialists working full-time on the development of the system. These included computer scientists, linguists and a small team of professional translators and terminologists, on loan from the federal government to flesh out the transfer dictionary. TAUM-Aviation was a classic second-generation system, decomposing the translation process into three major components: analysis, which produced a syntactico-semantic structure representing a certain formal understanding of the input sentence; transfer, which mapped that representation into an equivalent representation in the target language (TL); and generation, which transformed the TL structural representation into a properly ordered and inflected TL string. The linguistic operations performed by each of these components were described in hundreds, if not thousands of individual rules – either lexical entries or grammatical transformations – that were drafted one at a time by the linguists and translators, and formalised using high-level programming languages which the system's compilers could interpret and execute. The resulting system was prodigiously complex and not particularly robust.[5] Moreover, Aviation, like almost all NLP systems at the time, was uni-directional; most of the effort that went into hand-crafting the linguistic descriptions for the English-to-French system would not be able to be used in a system that translated in the opposite direction, i.e. from French to English. In the end, the cost of developing the system's elaborate grammars and detailed lexicons, which were absolutely necessary for the high-quality translation of this complex sublangauge, proved to be the project's undoing. An evaluation commissioned by Treasury Board (Gervais 1980) concluded somewhat controversially that producing translations with the Aviation system would be more expensive than the current cost of human translation. In 1981, the TAUM group was disbanded; and with its dissolution, Canada lost its position as a world leader in MT.

[4] Figures for 1990, according to (Grimaila 1992).

[5] During its development phase, the Aviation system was deliberately designed not to translate sentences that it could not fully analyse.

3. Machine translation now

For the first forty or so years of its history, machine translation was by and large the restricted purview of a small group of researchers and developers, and an even smaller group (or so it seemed) of MT system users. In those days, relatively few people had access to the mainframes on which MT systems ran and even if they did, few could afford the substantial licensing fees demanded by the vendors. The advent of the PC modified the economics of MT, making it possible for curious individuals to purchase the modestly priced translation software packages that began to appear in their local computer store. But even so, the number of people who used these packages on a regular basis for their on-going translation needs remained relatively small. This can be inferred from the fact that few of the companies that first offered PC-based MT systems are currently prospering – if they're still in business at all.

The Internet has radically changed all this. Thanks to the spectacular growth and the pervasiveness of the World Wide Web, more people today have access to and are actually using MT than ever before. This democratisation of MT is very recent and may conveniently be dated to AltaVista's BabelFish experiment, which was launched in late 1997. Today, there are numerous – I'm tempted to say dozens of – Web sites that offer free online translation services between a surprising range of languages.[6] True, many of these sites will only accept texts of limited length, being offered for the most part by vendors who eventually hope to sell their software package for real money. Nevertheless, the result of this MT-at-a-click phenomenon is undeniable: millions of people have actually experimented with machine translation now and so have at least some idea of what the current technology can and cannot do. And the figure of millions is by no means an exaggeration. One of the originators of the BabelFish site told an AMTA conference in 1998 that six months after its inauguration, the system was processing about a half a million translation requests… per day!

3.1 New users

The Internet has profoundly transformed the MT business in at least two distinct ways. On the one hand, it has facilitated general access to MT systems and MT service providers; and on the other, it has generated a tremendous new demand for translation, as millions of Internet surfers increasingly come across Web pages in languages they do not fully master. Of course, it is difficult to say with any certainty just who makes up this new and expanding population of MT users, but information on the BabelFish application does give us a clue. Again according to the above-mentioned source, the large majority of online translation requests handled by this system are for very short bits of text (less than three words on average), followed by Web pages. This does not at all correspond to the typical translation request submitted at the European Commission, where access to Systran is freely available to translators and bureaucrats alike and where the overwhelming majority of jobs submitted are full texts. So while more people may be using MT today than ever before, few of these new online users would appear to be professional translators.

The other significant characteristic of today's online MT has already been mentioned: it's free. Needless to say, the MT industry has attempted to capitalise on, i.e. to make money from,

[6] For a comprehensive list of free online translation sites, see http://www.eamt.org/resources/index.html .

the enormous new demand for translation, which, owing to its sheer volume and transitory nature, cannot possibly be satisfied by traditional translation services. Recently, for example, companies like Mendez (formerly a branch of Lernout & Hauspie), Uniscape and Canada's own Alis Technologies have set up a kind of one-stop translation portal on their Web sites, offering a range of services, from raw or revised machine translations to polished human translations. It is still too early to say whether this new way of marketing MT will turn out to be a profitable. Are individuals and corporations really prepared to pay for gist-quality translations over the Web? Will they entrust their high-quality translation needs to anonymous translators, rather than dealing with local providers whom they know and have confidence in? Be that as it may, free MT on the Internet is certainly rendering a great public service. Moreover, in making foreign-language Web pages (at least partially) comprehensible, it may actually be whetting people's appetite for fully comprehensible, polished translations. Far from posing a threat to translators, MT on the Web may actually result in a substantial increase in the demand for human translation.

3.2 New development methods

Between 1990 and 1994, the Defence Advanced Research Programs Agency (DARPA) of the United States government sponsored a large-scale competitive development program in machine translation. Intended to significantly advance the state of the art, the program pitted three different MT approaches against each other: a novel statistics-based approach championed by researchers at IBM's T.J. Watson Research Centre, who developed a system called Candide; a rule-based interlingual approach developed by well-known experts at several American universities, whose system was called Pangloss; and a hybrid approach combining statistical and rule-based methods, which was developed by a team that had made its mark in commercial speech recognition. The DARPA program involved frequent evaluations, or "bake-offs", in which the three systems were required to produce translations of unseen texts. These translations were scored along various parameters by a separate, well-funded team that elaborated a sophisticated evaluation methodology. The performance of the three research prototypes was also compared to that of a few well-known commercial MT systems.

The DARPA program proved to be a resounding success; indeed, one could say that it changed the face of machine translation forever. The IBM team vaunted the fact that their Candide system made no use whatsoever of explicit (or declarative) linguistic knowledge, but rather induced its translation and target language models directly from large corpora of raw text. Inspired by techniques that had previously proven successful in speech recognition, they trained a word correspondence model on three million aligned sentences taken from the Canadian Hansard; in their French-to-English system, this was used to replace the source words in a new text with their most likely English translations. These forms were then reordered by a probabilistic target language model, which had been trained on a large corpus from the Wall Street Journal.[7] The training of Candide required huge machines and the resulting system was totally opaque and impossibly slow; and yet contrary to all expectations, it outperformed the other two research systems and came close to matching the performance of Systran, the leading commercial system. These results sent tremors through the entire MT and NLP communities. As if to rub salt into their opponents' wounded pride, members of the IBM team revived a celebrated

[7] This is something of a oversimplification; in particular, some of the target language ordering information could derive from the alignments of the translation model, which would then be refined by the language model.

quotation from Fred Jelinek, a pioneer in automatic speech recognition (SR): "Every time I fire a linguist, my [system's] performance goes up." Statistical MT had arrived.

4. The new empirical paradigm in NLP [8]

As alluded to above, the emergence of the new data-driven, or empirical methods as they are often called, first began in speech recognition and gradually extended to all the other branches of natural language processing. Faced with perennially disappointing results, Jelinek and his fellow SR researchers at IBM finally decided to abandon the linguistically motivated models which were prevalent at the time, in favour of cruder but more robust noisy channel models first proposed by Shannon and Weaver in information theory back in the 40's and 50's. The new SR systems employed ngrams and Hidden Markov Models to estimate and encode the relative probabilities of sound and word sequences; and though Chomsky had convincingly demonstrated the inadequacy of such formalisms as models of *human* linguistic competence, they consistently outperformed the linguistically motivated systems on the practical engineering task of speech recognition.

The explosion of the Internet was another important factor that contributed to the resurgence of 1950's style empiricism in NLP. Staggering quantities of textual data suddenly became available in the late 1980's, through which millions of Internet users urgently needed to search. The symbolic approaches previously employed in information retrieval (IR) simply could not cope with document collections on this scale. Instead of elegant applications in self-enclosed domains, computational linguists and computer scientists began developing IR systems based on rudimentary but robust techniques that work relatively well over unrestricted text in domains as vast as the Internet. These eventually gave rise to the search engines, like Yahoo and Google, that we now use every day.

For a number of years, a basic scission persisted in the NLP world between those in SR and IR, who relied essentially on data-driven techniques, and the old-school computational linguists, who continued to employ their rationalist techniques and symbolic rule formalisms. Each had their own journals, organised their own conferences, and there was very little exchange between them. But eventually, the latter too began applying empirical techniques borrowed from SR and IR to the solution of well-defined, generic problems in NLP, like part-of-speech tagging, word sense disambiguation and parallel text alignment.[9] At first, it was generally thought that the trade-off would be between high quality and robustness. Because they are based on a deeper analysis of the linguistic phenomena, the symbolic, rule-based approaches are capable of achieving high quality output – when they work. The stochastic approaches, on the other hand, tend to consider more superficial linguistic phenomena, but are far less likely to crash in the face of unexpected types of input. It turns out, however, that the empirically-based methods have gradually supplanted classical rule-based methods in practically all NLP domains. Even on the traditionally difficult task of parsing, on which so many linguists have worked for so long, statistical parsers can now perform at least as well, if not better, than their hand-crafted counterparts.[10] Machine learning techniques now completely dominate computational linguistics; to convince oneself, one need only consult the program of any recent ACL conference. Virtually

[8] For the information in this section, I have drawn heavily on (Hovy et al. 1999), particularly chapters 4 and 6.
[9] These are generic in the sense that the results can be applied to various problems, as we will see below.
[10] For example, see (Charniak 2000) and the references cited there.

all of the NLP systems developed today are at least partially self-organising, in the sense that they acquire their linguistic knowledge (semi-) automatically, through training on large corpora of raw, annotated or sometimes even structured data, like that found in machine readable dictionaries or thesauri. Does this mean that traditional computational linguists will soon be out of work? (Recall Jelinek's remark quoted above.) Not necessarily. These new self-organising systems still need to be told what is it that they must learn from the data; if they use annotated training data, that data must be annotated according to some theory. Part-of-speech taggers, for example, need to be trained on texts that linguists have annotated with the correct part-of-speech for each word; before they can do that, of course, the linguists must first define the tag set and its level of granularity – decisions which should in principle be based on some linguistic theory. Similarly for the statistical parsers mentioned above. Most of the work in this area has been based on the Penn Tree Bank, whose labelled bracketings were produced and verified by linguists over years of painstaking effort and no small expense. Indeed, if there is one major drawback to the new empirical paradigm in NLP, it is this: the effort and cost often required to develop large, ideally reusable amounts of training data. The pay-offs, on the other hand, in terms of practical, robust, real-world applications have certainly been substantial.

5. Recent research in machine translation

In 1992, the MT group at the CITI of which I was a member organised a conference on the theme of "Empiricist versus Rationalist Methods in MT". The timing couldn't have been more propitious, for the conference was held at the height of DARPA's MT initiative. TMI-92 attracted most of the leading researchers in MT at the time, and the discussions following the papers and the panels were highly animated. The conference also featured keynote addresses by Yorick Wilks, a member of the Pangloss consortium, and Robert Mercer, a leader of the IBM Candide team. Mercer was particularly provocative in his remarks, stating at one point that "rationalist methods in MT will be on the scrap heap five years from now."

A conference on the same theme would be a flop today, for there is no longer any controversy over the use of data-driven methods in MT. That debate is dead, and it would be difficult to find anyone still prepared to defend a purely rationalist approach to MT development.[11] Mercer's provocative remark has proven to be true. Although there still exist longstanding MT projects and systems that use rationalist methods to update their rule base, few if any new projects now rely on traditional labour-intensive methods, like those that were employed on the TAUM-Aviation project, to build up their dictionaries or grammars – if indeed their systems still incorporate such distinct linguistic components. One undeniable and extremely significant contribution of the new empirical paradigm in NLP has been to make available alternative techniques that are much less costly in terms of human labour. (More on this below.)

MT research, moreover, is definitely on the upswing. We are finally beginning to see an encouraging number of new MT projects, not only in Japan, where their export-based economy has long made translation a national priority, but also in Europe, where the number of official EC languages has created a translation problem of staggering proportions, and even in the United States.[12] Given the pressing multilingual demands of growing globalisation, large corporations

[11] If by that one means a predominant reliance on linguists' intuitions to develop a system's lexicons and grammars.
[12] Though sadly, not in Canada.

like Microsoft, AT&T, Sony and IBM (to name just a few) are now investing substantial sums in MT research. Why now? What has changed in recent years to explain this renewal of interest and investment in MT? For one thing, people are finally beginning to realise the full extent of the looming crisis in the translation industry. The demand for high-quality translation – whether it be of user manuals, promotional literature, or the reported proceedings of international bodies – is growing by leaps and bounds, and already surpasses the capacity of the translation profession to meet it; and this tendency shows no sign of abating. At the same time, the quality of the translations produced by existing MT technology is still not generally sufficient to allow for cost-effective post-editing. Hence, there is tremendous market pull today. In plain terms, more businesses and international organisations than ever before urgently need a solution to the problem that is often designated as the translation bottleneck. For anyone who can come up with even a partial solution – a technology that will allow even modest productivity gains – the rewards will be tremendous. This is the primary factor driving the recent renewal of interest in MT research. A second, ancillary factor is the hope held out by the new empirical methods in NLP that substantial progress, even if no spectacular breakthrough, may at last be possible.

Let me now briefly mention some of the major trends in recent MT research and development.

5.1 Controlled language

The first of these is not really in MT at all, but may actually be viewed as betraying a certain scepticism regarding MT. I'm referring to the use of controlled languages (CL) as a way of simplifying the task for MT.[13] Under this approach, linguistic restrictions are imposed on the drafting of the source text so as to eliminate certain difficulties – e.g. terminological polysemy or various syntactic constructions, such as co-ordination or noun compounds – which are known to pose a problem for machine translation. Simplifying the source texts in this way is intended to ensure better quality MT output, minimising the need for post-editing and reducing overall publication costs. It turns out, however, that introducing CL is more difficult than one might imagine. For one thing, those people drafting the source texts do not take well to having their prose artificially constrained. Early experiments in CL merely provided technical writers with guidelines which they were encouraged to obey but often blithely ignored. The proponents of CL soon realised that the only way to enforce the linguistic restrictions they wanted to impose was to develop automatic editing programs that would flag constructions and vocabulary that did not conform to the CL specifications and suggest alternate reformulations. This kind of interactive editor is now a central component of several large-scale CL projects, including the one that CMU developed for Caterpillar in the 1990's; and even so, Caterpillar has had to set up extensive training programs for its technical writers. In fact, to my knowledge, there have been no published reports on the overall cost-benefits of this high-profile project; although this has not prevented other large corporations with similar translation needs, e.g. GM, from launching their own CL projects recently. As should be obvious, this kind of solution is only applicable to a narrow range of translation situations. In particular, controlled language lends itself best to the multi-target translation of technical documentation, where there is a certain tolerance for short,

[13] Nor is it altogether new. Xerox, for example, has reportedly been using CL for years, in conjunction with Systran, to translate its technical manuals. What *is* new, however, is the recent upsurge of interest in CL.

simple, repetitive sentences. Moreover, the introduction of CL requires tight co-ordination between those who draft the documentation and those who translate it.

5.2 Example-based machine translation

Returning now to MT research proper, two novel approaches emerged over the last decade which have proven to be dominant. One, statistical MT (SMT), has already been discussed above. The other, generally referred to as Memory-based or, more commonly, Example-based machine translation (EBMT), was first proposed by (Nagao 1984) and has been particularly popular in Japan. Recall that classic MT systems decompose the translation process into three components: analysis, transfer and generation. Nagao argued that "the detailed analysis of the source language sentence is of no use for the translation between languages of completely different structure like English and Japanese." (p.179) Moreover, such structural analyses are highly theory-dependent, and unfortunately linguistic theories can and often do change; whereas raw language data change very little over time. His proposal therefore was to construct an MT system which would be much less dependent on the analyses provided by linguistic theory, but relied instead on primary unanalysed data, in the form of paired examples of translated sentences. Rather than decomposing the input sentence into a structural representation,[14] and then mapping that representation into a target language equivalent via complex transfer rules, an EBMT system would search its translation database for an identical (string) match of the new input; and if one was found, it would retrieve the associated TL translation. This, of course, is now very familiar and has come to be known as translation memory (TM), or repetitions processing.[15] But Nagao's proposal went further: if no exact match was found in the database, the system would try to locate source sentences that were similar to the new input sentence, using a thesaurus to measure the semantic distance between words.[16] Moreover, if several partially similar sentences were retrieved from the database, the system would recombine the appropriate fragments to produce a coherent target output. Like TM, one of the principal attractions of EBMT is that it relies on previous human translations; this is seen as a guarantee of a certain level of quality. Unlike TM, however, EBMT aims for full-fledged automatic translation, and unfortunately Nagao may have underestimated the difficulty of the program he was advancing. In particular, there would seem to be an inherent tension in this approach between the drive to enlarge the database of past translations and the difficulty this creates for evaluating competing matches. Most researchers in EBMT soon abandoned the simplistic string-based matching strategy in favour of storing structured sub-trees with cross-language correspondences explicitly indicated at multiple levels, in contradiction to Nagao's original desire to rely as far as possible on the unanalysed data. And while significant energy and resources have been invested in this intriguing idea, I think it is fair to say that thus far the results in terms of operational systems have been somewhat disappointing. Recent efforts have sought to develop hybrid systems that combine EBMT techniques with other architectures, or even restrict their application to certain problematic constructions. For a comprehensive overview of the work done in EBMT, see (Somers 1999).

---

[14] Alternatively, one could respond to Nagao's objection by arguing that the structural analysis of typologically different languages must abstract away from their surface differences and aim for deeper, more semantic representations.

[15] For a critique of current TM technology, see (Macklovitch and Russell 2000).

[16] This is very different from the character-based fuzzy matching measures employed by many TM systems.

5.2 Spoken language translation

One final trend that cannot go unmentioned is spoken language translation (SLT). There were two very large projects in SLT over the past decade, Verbmobil and C-Star, the latter, an international effort, is still ongoing. SLT is a remarkably ambitious undertaking, combining all the difficulties of speech recognition with many of the traditional problems of MT. These problems are actually exacerbated by the fact that spoken language contains many "ill-formed" utterances, i.e. sentence fragments (which frequently must be pragmatically interpreted), self-corrections and uhh… hesitations. Hence, SLT cannot possibly be tackled by simply coupling speech recognition and voice synthesis modules to a classic, text-based MT system. Specialised MT modules are required that can handle such input and respond in near real-time. In fact, many of the exciting advances in statistical MT in recent years have been developed to meet the needs of SLT. Still, even the most evolved research prototypes in SLT today limit themselves to collaborative dialogues in very restricted domains, such as making travel reservations.

6. Recent developments in machine-aided translation

Although Candide's performance in the DARPA bake-offs sent shock waves through the MT community, it must be recalled that the statistically-based system did not actually outperform Systran, the oldest and best-known commercial system, with its enormous hand-crafted dictionaries painstakingly elaborated over decades. In fact, Candide seems to have peaked about half-way through the DARPA program, leading the IBM team to introduce a modest measure of linguistic pre-processing in order to strengthen the distributional regularities which the system would acquire from the corpora, c.f. (Brown et al. 1992). What exactly was responsible for the performance ceiling that seems to have stymied Candide? The answer is not all obvious, but the following is the view of (Chanod et al. 1999): "unlike speech recognition, translation cannot operate adequately at the word level, but must involve more abstract constructs such as syntax." (p4)

Be that as it may, Candide undeniably revolutionised MT development. The project's principal legacy has been methodological: thanks to the data-driven techniques first elaborated by the IBM group, MT systems can now be developed in a fraction of the time and at a fraction of the cost of the traditional rule-based systems. This was convincingly illustrated recently by the challenge which the members of the 1999 Johns Hopkins summer workshop on statistical MT set themselves: to develop an MT system for a new language pair within a single day – and they succeeded; see (Al-Onaizan et al. 1999). Of course, this tells us nothing about the quality of the translations produced by the resulting Chinese-English system. Still, even former sceptics now grudgingly admit that new MT systems can be developed using statistical methods which are at least comparable in their performance to the levels of the older systems, which were so costly to develop. In itself, this is a remarkable achievement, with enormous repercussions for many of the world's so-called minor languages.

Another notable feature of recent research in MT is that it focuses almost exclusively on fully automatic systems. Will these new projects eventually overcome the above-mentioned performance ceiling,[17] yielding systems that can significantly outperform their classical

[17] Interestingly, some of the best recent work in SMT attempts to go beyond word-for-word substitution, in order to model phrasal translation correspondences; c.f. (Och et al. 1999)

counterparts? That is difficult to say, but there is at least one factor which should caution us against excessive optimism. The hurdle posed by the extra-linguistic factors – all the knowledge that is necessary for high-quality translation but is not actually present in the text itself – remains as daunting for SMT as it was for classical rule-based MT. On the other hand, many of the same empirical techniques that have recently been developed for fully automatic MT systems can also be profitably applied to less difficult tasks, such as lexicon building and automatic term extraction, as well as to the development of other sorts of translator support tools. Indeed, this has been the focus of our laboratory at Université de Montréal: exploiting the new probabilistic techniques in NLP to develop innovative aids for human translators. I would now like illustrate some of this work, beginning with applications based on very simple techniques and showing how they can still be made to yield considerable assistance for human translators and terminologists, before moving on to describe a much more ambitious project in interactive MT.

6.1 Sentence alignment for bilingual concordancing

In section 3.2 above, we mentioned that the Candide system was trained on three million aligned sentence pairs taken from the Canadian Hansard. Although I'm convinced it was done quite unwittingly, the Canadian government has in fact made an enormous contribution to recent research in computational linguistics by making available the translated proceedings of this country's parliamentary debates. For the purposes of statistical MT, however, there remains a problem with the Hansard corpus as it is found on the House of Commons' Web site: the two language versions are published in separate files. In order to be able to extract the translation knowledge that is latent in these texts, it is first necessary to align them, i.e. to explicitly link the corresponding segments in the two texts. In principle, this can be done at different levels of granularity; however, as one descends the linguistic hierarchy, from the level of the paragraph to that of the sentence, the phrase, the word and the morpheme, the number of one-to-one correspondences tends to diminish, making the task of automatic alignment more difficult. Over the last decade, considerable effort was invested into developing automatic alignment algorithms; for a good overview of the state of the art, see (Véronis and Langlais 2000). The upshot is that we are now able to automatically align vast quantities of parallel texts with a high degree of accuracy at the sentence level, thereby creating what Brian Harris first called bi-textual databases.

One straightforward way of exploiting such bi-textual databases is via a bilingual concordancer, like the RALI's *TransSearch* system. The basic idea is very simple. The user submits a translation problem to the system in the form of a query, which may be a single word, a phrase, an expression or even an entire sentence. The system searches its database for all occurrences of that query and displays each occurrence within its full sentential context. Alongside each occurrence, moreover, the system also displays the *translation* of that sentence. It can do this because each sentence in the database has been explicitly linked to its translation through the automatic alignment procedure. Hence, each retrieved occurrence brings with it a tentative solution to the original problem, previously proposed by another translator. While we call *TransSearch* a bilingual concordancer, our users tend to view the system as an enormous, virtual bilingual dictionary of examples. And the database *is* enormous: *TSrali.com*, the new commercial version of *TransSearch*, includes all the Canadian parliamentary debates from 1986 to the present – over 180 million words of translation – as well a sizeable collection of decisions from the Supreme Court of Canada, the Federal Court and the Tax Court of Canada. A snapshot

of the Web-based query interface is given in Figure 1, which appears in the Appendix to this paper. Presumably, what the user wants to know in this example is how to translate the expression "cut to the bone". Notice, however, that the ellipsis in the query will allow for other intervening material between the first and last words. Notice too the plus sign appended to "cut": this will allow for the retrieval not just of the base form of the verb but also all its inflected forms. Once a query is submitted, *TransSearch* displays the results on a separate Web page, listing ten results per page; c.f. Figure 2 in the Appendix. If the user needs to see more context for any translation pair, he or she need only click on the number in the left-hand margin. For more detailed information on the query language and the internal structure of the *TransSearch* system, see (Macklovitch et al. 2000).

6.2 Part-of-speech tagging for automatic term extraction

Part-of-speech tagging was mentioned above as one of the generic tasks that computational linguists tackled at the beginning of the last decade, using similar techniques to those that had proven successful in speech recognition and information retrieval. What exactly is meant by part-of-speech (POS) tagging? The term is normally employed to refer to (the output of) a program that assigns a grammatical category, or tag, to every word-form in a text. What makes the problem difficult, of course, is that many of the words in any natural language are categorically ambiguous, i.e. they can belong to more than one POS. Consider, for example, the form "light" in English, which can be either an adjective, a noun or a verb; and similarly, the form "ferme" in French. A *stochastic* part-of-speech program calculates the most *likely* POS assignment for each word in a text, on the basis of two sets of probabilities: first, each form's lexical probability, i.e. the relative likelihood that a given word will appear as category x versus category y in any text; and second, its contextual probability, i.e. the likelihood of observing category z in a text, given the category of a small number of words that precede it. Now standard dictionaries do not generally inform us about lexical probabilities; they merely enumerate all the categorical possibilities for a given word. Thus, if we were to look up in any English dictionary the words in the sentence “I see a bird”, we would be somewhat surprised to discover that each form is categorically ambiguous.[18] In particular, “see”, in addition to being a verb, can also be a noun, referring to a bishop's area of jurisdiction; and “bird” can also be an action, referring to something that bird-watchers do. *We* know that both of these uses are quite rare, of course; but for a long time, few if any natural language parsers had access to this kind of frequency information, and so it was common for them to produce what humans felt were the oddest of multiple readings.

One of the first and best-known POS taggers was developed by Ken Church, see (Church 1988). It was trained on the one-million word Brown University corpus, whose tags had been manually and laboriously assigned over a period of many years. Church's program incorporated information on the relative frequency of each form with regard to its different POS assignments in the Brown corpus; it learned, for example, that “see” occurs 771 times as a verb in this reference corpus and only once as a noun. The program also had access to statistics on trigrams, or more exactly three POS tag sequences. Now linguists have long known that certain dependencies exist in natural language which make it impossible to reliably determine the category of a given word-form solely in terms of the two categories that precede it. Church never

[18] This example is borrowed from (Church 1988).

denied this; instead, he argued that such phenomena turn out to be relatively infrequent in real-life texts. And he demonstrated this convincingly by showing that his tagger could consistently achieve accuracy rates of between 95-99%.

Stochastic POS taggers now form part of many NLP applications, from speech synthesis to sophisticated database query systems. Here, I would like to illustrate a somewhat simpler application of this now standard technology, in the form of a monolingual term extraction program developed by the RALI called *ExTerm*. *ExTerm* works roughly as follows: Given a preferably technical text as input, the program first assigns a POS tag to each word in the text, in much the same manner as Church's program described above. It then lemmatises all the inflected forms, reducing each word to its base form. Once this is done, *ExTerm* searches the tag sequences for patterns that correspond to its definitions of French or English multi-word terms – for example, two common nouns preceded by an adjective in English; or a noun followed by a preposition plus a noun in French. Finally, it extracts these candidate terms and sorts them in order of descending frequency; any multi-word sequence that occurs at least twice in the text is proposed to the user as potential term. An example of *ExTerm* output is provided in Figure 3 in the Appendix.

I will limit myself to a few brief remarks on *ExTerm*, since the topic of automatic term extraction will be treated more fully by other people at this Symposium. Despite the fact that *ExTerm* is based on very simple technology, the program performs surprisingly well. That is, the most frequently occurring candidate terms proposed at the top of the list almost invariably turn out to be bona fide terms. The program's performance is less good, however, when measured in terms of recall; i.e. there may be many terms in a text which the program does not identify, most notably all single-word terms. *ExTerm* exploits an interesting observation about multi-word terms in English that was first made by (Justeson and Katz 1993); to wit, that any multi-word sequence of adjectives and nouns ending in a noun which reoccurs verbatim more than a certain number of times is very likely to a term.[19] Unfortunately, this does not tell us how to reliably distinguish single-word terms from other nouns that form part of the general vocabulary of the language – no easy task, even for a human terminologist. Nevertheless, programs like *ExTerm* can be used to relieve some of the tedium of manually scanning texts for terms, thereby increasing the productivity of terminologists. All things being equal, moreover, the better the POS tagger, the less noise there will be in the list of candidate terms, and hence the greater the potential impact on productivity.

6.3 Probabilistic translation models for the automatic extraction of bilingual lexicons

Returning again to Candide, recall that IBM's SMT system was actually comprised of two distinct components: a translation model and a target language model. Although they were combined in Candide to produce fully automatic translations, each of these components can in fact be employed for various other purposes. For example, the RALI uses a probabilistic language model much like that of Candide in its automatic French reaccentuation system, *Réacc*. It is this language model which selects the most likely accented form for each word, given the two words that precede it. For a detailed account of the *Réacc* system, see (Simard 1998).

[19] Similar POS sequences that are non-terminological descriptive noun phrases, on the other hand, are much less likely to be repeated verbatim and tend to be reduced to their head noun.

Probabilistic translation models can also be employed for purposes other than fully automatic MT, most obviously to help lexicographers extend the coverage of various sorts of dictionaries, or to help human translators construct their own domain-specific glossaries. Indeed, one of the principal attractions of these statistical techniques is that they lend themselves easily to new language pairs or new technical domains, without requiring supplementary linguistic resources or highly skilled specialists. Once the programs have been elaborated, all that is required are the appropriate parallel texts in sufficient quantity. At the RALI, for example, Professor Jian-Yun Nie has shown how parallel texts, mined from the Web, can be used to develop translation models for various language pairs, these models then being applied to the task of cross-language information retrieval with very encouraging results; c.f. (Nie et al. 1999).

I will not attempt to describe in any detail how the algorithms that underlie these statistical translation models function; they are exceedingly complex and surpass my limited understanding of statistics. (The interested reader may consult (Knight 1999) for a practical introduction.) Basically, however, all such models rely on parallel corpora to compute statistics on the frequency with which source and target language words co-occur in aligned segments. The resulting models are probabilistic in that they furnish no absolute translation correspondences, only more or less probable ones. In other words, every target word in the training corpus is viewed as being a potential translation of each source word; all that distinguishes the small set of intuitively plausible translations from all the entirely outlandish ones is that the former should have much higher probability scores. In their seminal work on the mathematics of SMT, (Brown et al. 1993) define a series of five increasingly complex translation models that take into account an expanding range of factors in establishing the alignments between words; but central to all these models is the probabilistic bilingual lexicon, i.e. the table that lists the target equivalents for each source word, along with their probabilities. (Melamed 1998) describes several ways of refining the basic methodology so that the extracted translation lexicons become sufficiently accurate to allow for their integration into the workflow of conventional MT system developers. But as Melamed points out, these methods are designed to assist lexicographers and not replace them, for while they can reliably provide the possible translations of a given source word, they do not make explicit the various contextual factors which determine when a given equivalent is appropriate. Then again, neither do standard bilingual dictionaries or term banks.

Another significant limitation of the statistical models developed by (Brown et al. 1993) is that they are essentially limited to word-for-word equivalences; they don't explicitly account for translations between multi-word phrases and single words, or between phrases and phrases. Needless to say, the IBM group understood full well that this is not true of translation in general; if their models fail to allow for what they call the "general alignments" of phrases to phrases, it was essentially for reasons of computational tractability. At the RALI, Philippe Langlais has recently done some interesting work designed to overcome this limitation on statistical translation models. Beginning with a large bi-text of aligned sentences, he first applies the POS tagging technology described in the preceding section to identify what he calls salient phrasal units in each language. Each such unit is then graphically fused so that it is like a single word in the text. Next, he trains a new translation model on this bi-text, which calculates the correspondences between units across the languages. Fusing together the individual words in the units, however, aggravates what is known as the sparse data problem: when there are too few occurrences of source and target language expressions, the accuracy of the calculated correspondences may suffer. To counter this problem and increase the accuracy of the bilingual

mapping, Langlais introduces a number of linguistic filters; c.f. (Langlais et al. 2000) for details. A sample of the extracted bilingual pairs that Langlais has obtained from a Hansard corpus appears as Figure 4 in the Appendix. The list still contains a few bugs which can undoubtedly be corrected; but overall, the quality (or precision) of the associated units is very good.

The usefulness of this kind of technology should be quite obvious. If you have sufficient quantities of parallel text, these programs can automatically extract a bilingual lexicon in a fraction of the time and effort it would take to create the same lexicon by hand. (Melamed 1998) claims that his programs can achieve over 90% accuracy and 90% coverage. Of course, all such methods involve a trade-off between recall and precision. In Langlais' work, the emphasis has been placed on precision, with a view to limiting the time a human would have to spend rejecting false equivalents in a validation cycle.

6.4 Interactive machine translation

As mentioned above, almost all the new projects employing data-driven techniques in MT aim at developing fully automatic systems. Given the current state of the art, however, there is a price to pay for this: full automation can only be obtained by sacrificing either translation quality or the generality of the text types the system can handle. What if one is not prepared to forego either quality or generality? In this section, I want to briefly describe an ambitious research project in interactive machine translation (IMT) called *TransType*, which the RALI has been working on for the last four years.

IMT is actually a fairly old idea; see (Kay 1973) for one of the earliest proposals. Unlike the standard model of MT collaboration in which the machine first translates and the human then intervenes to post-edit and correct, in IMT the machine and the human iteratively interact to jointly produce the target text; hence, there is no post-editing in the standard sense. In almost all previous IMT systems, the locus of this interaction has been the source text: typically, the system requests the human's assistance in disambiguating various aspects of its meaning or structure, even though this is not something that translators (as opposed to linguists) are generally trained to do, at least not explicitly or formally. In *TransType*, on the other hand, the locus of the interaction is the target text itself. The system can be conceived as a kind of accelerated typewriter designed to speed up the translator's work by automatically completing the target unit that s/he has begun to key in. It does this by reconciling the "prefix" of the translation unit that the human has already typed with the translations predicted by its own stochastic translation and language models. In *TransType*, these predictions are recalculated in real time, with each new character the user inserts.[20] By exploiting these statistically derived predictions, (Foster et al. 1997) show how *TransType* can predict up to 70% of the characters in the translator's intended target text. Figure 5 in the Appendix provides a snapshot of the interface of the current *TransType* prototype.

One of the advantages of IMT is that it allows the human to retain complete control of the translation process. *TransType's* role is supportive: it assists the translator by speeding up the entry of the target text and by suggesting translations when the translator is stumped or otherwise runs out of inspiration. However, the translator is always free to accept or to ignore the system's

[20] Advances in computer hardware and more efficient statistical models now allow for more rapid SMT systems than the original Candide system.

suggestions; in the worst case, s/he simply continues to key in the intended target text, as though the system were no more than a text editor. But this should rarely happen, since *TransType* will frequently save time and avoid transcription errors by automatically reinserting difficult proper names and troublesome numerical expressions. Another major advantage of *TransType* is that it is an adaptive MT system. The translator may decide to change his or her mind, delete a segment of target text and begin typing a new translation; *TransType* verifies the prefix of the new text and modifies its suggested completions accordingly. Unlike classical MT systems, *TransType* is not designed to generate a single correct translation for each source sentence, but a set of more or less probable translations.

I do not want to suggest that this new approach to IMT is entirely without problems. On the contrary, *TransType* raises non-negligible ergonomic issues that have not yet been adequately resolved. In particular, how can the system best suggest its completions without disrupting the translator's train of thought? Moreover, the completions suggested by the current prototype are limited to single words and some fixed expressions. We would obviously like to extend these completions so that the system can predict longer units of target text, perhaps even complete sentences. Both of these problems will be addressed in a new international research project based on the RALI's *TransType* system, a project that will add a vocal interface to the system and two new target languages (German and Spanish) to its repertoire. For an update on the technology underlying *TransType*, see (Foster 2000).

7. Conclusion

Returning now to the issues raised in the introduction of this paper, I hope to have demonstrated that we have indeed witnessed a fundamental paradigm shift in the field of natural language processing over the last fifteen years. The types of questions now asked by computational linguists and computer scientists interested in language, the problems they consider important, above all the experiments they perform to investigate those problems – all this has changed enormously in recent years, to the point that it does not seem inappropriate to speak of a paradigm shift, in the sense first proposed by (Kuhn 1962). Furthermore, this paradigm shift has had far-reaching consequences on the various sub-domains of NLP, including applied research and development in machine and machine-aided translation. Where twenty-five years ago, development work in MT was extremely *labour-intensive*, requiring large numbers of highly-skilled professionals, today the same kind of work has become *data-intensive*, requiring powerful computers, very large corpora and a smaller number of specialists with very different training. The new empirical methods in NLP, and in particular the new machine-learning techniques, have revolutionised our field. Overall, I would say this is a welcome development; one could even call it progress. For while the cost of skilled human labour has increased significantly over this period, the drop in the cost of computing power has been even more spectacular; and more electronic text is available today than ever before.

Does this mean that we are significantly closer today to FAHQT, or fully automatic high quality translation of unrestricted texts, that Holy Grail which Bar-Hillel first defined in the early 1950's? This is as much a judgement call as a scientific question. Here is my view of the issue. While there has undeniably been substantial progress in the general quality of the machine

translations produced by today's best systems,[21] to the point that MT for gisting purposes is now altogether viable for related language pairs, much distance nevertheless remains to be covered before we can comfortably combine the terms high-quality and MT within in the same phrase. As I stated above, there will always be an irreducible core of translation problems in every text of nominal size whose solution requires recourse to real-world, or extra-linguistic knowledge. As yet, we have no way of incorporating such knowledge into our statistical models, and until we do, I doubt that the new empirical methods will be able to perform miracles. However, the principal impact of the new paradigm has not so much been on translation quality as on the cost of system development. Because the programs that analyse large corpora and produce these statistical models are entirely reusable, we are now able to produce new MT systems in a fraction of the time and cost that was necessary twenty-five years ago. One should not underestimate the importance of this for many of the world's so-called minor languages.[22]

Overall, I would say that the prospects for the future appear brighter today than they have been in a long time. In fact, I am rather optimistic that the new research paradigm in NLP will actually be able to help attenuate the current crisis in the translation industry. There have been several important changes which would seem to warrant such optimism. For one, the general public now has a much better understanding of the difficulties of translation and the limits of automation. In large measure, we have the Internet to thank for this, and it is a good thing. Thanks to MT's increased exposure, people are beginning to recognise that there are different types of translation needs, and that for certain of these, fully automatic MT can often do an adequate job, e.g. translation for assimilation purposes. As a result, the special skills of highly qualified human translators can be reserved for more important jobs. Moreover, the quality of MT output should continue to improve over the coming years, via the integration of proven NLP techniques such as named-entity recognition. When added to MT systems, this technology will help avoid the embarrassing errors produced when the system mistakenly translates proper names.

Finally, I have also tried to demonstrate that the new empirical techniques in NLP offer the possibility of significantly increasing the productivity of human translators through the development of exciting new support tools and translator aids. While we may not be on the verge of replacing human translators with translating robots, we should be able to use these scarce and precious human resources more strategically and effectively. And that too would certainly constitute non-negligible progress. So what does the future hold in store for the translation profession? I would say: work, work and more work… Hopefully, the new research paradigm in NLP will be able to help translators better cope with their ever-increasing workload.

[21] (Beaven 1998) provides authentic examples of translation output from EC Systran for different language pairs, where the same source sentences have been resubmitted to the system ten years later. As the author points out, the sampled sentences have no statistical significance; nevertheless, the difference in quality between the two sets of output is quite striking.

[22] Even under the new paradigm, however, one important prerequisite remains, and that is the availability of large parallel corpora.

8. Acknowledgements

I gratefully acknowledge the scientific contributions and unflagging support of all the members of the RALI, past and present. Special thanks to Graham Russell and George Foster for helpful comments on a preliminary version of this paper. I also want to thank Pierre Isabelle, who was instrumental in forming this team and whose short article, (Isabelle 1993), served as the inspiration for this one.

9. References

Al-Onaizan, Y., J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F-J. Och, D. Purdy, N. Smith, D. Yarowsky. (1999) *Statistical Machine Translation*, Final Report, JHU Workshop.

Bar Hillel, Y. (1951) *The State of Machine Translation in 1951*, in **American Documentation**, vol. 2, pp.229-237.

Beaven, John. (1998) *MT: 10 Years of Development*, in **Terminologie et Traduction**, La revue des services linguistiques des institutions européennes, Luxembourg, pp.242-256.

Brown, P., S. Della Pietra, V. Della Pietra, J. Lafferty, R. Mercer. (1992) *Analysis, Statistical Transfer and Synthesis in Machine Translation*, in the **Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation** (TMI-92), Montreal, Canada, pp.83-100.

Brown, P., S. Della Pietra, V. Della Pietra, R. Mercer. (1993) *The Mathematics of Statistical Machine Translation: Parameter Estimation*, **Computational Linguistics**, vol. 19, pp.263-311.

Chanod, J-P., J. Hobbs, E. Hovy, F. Jelinek, M. Rajman. (1999) *Methods and Techniques of Processing*, appears as chapter 6 in (Hovy et al. 1999).

Charniak, Eugene. (2000) *A Maximum-Entropy-Inspired Parser*, in the **Proceedings of the first Meeting of the North American Chapter of the Association for Computational Linguistics** (NAACL-2000), Seattle, Washington, pp.132-139.

Church, Kenneth W. (1988) *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text*, in the **Proceedings of the Second Conference on Applied Natural Language Processing**, Austin, Texas, pp.136-143.

Foster, G., P. Isabelle, P. Plamondon. (1997) *Target-Text Mediated Interactive Machine Translation*, **Machine Translation**, vol.12:1-2, pp.175-194.

Foster, George. (2000) *A Maximum Entropy/Minimum Divergence Translation Model*, in the **Proceedings of the ACL-2000 Conference**.

Gervais, Antoni. (1980) **Évaluation du système pilote de traduction automatique TAUM-Aviation**, Secrétariat d'État, Ottawa.

Grimaila, Annette. (1992) *Made to measure solutions*, in J. Newton (ed.) **Computers in Translation: A Practical Appraisal**, Routledge, London, pp.33-45.

Hovy, E., N. Ide, R. Frederking, J. Mariani, A. Zampolli (eds.) (1999) **Multilingual Information Management: Current Levels and Future Abilities**, commissioned by the US National Science Foundation, available online at http://www.cs.cmu.edu/~ref/mlim/ .

Isabelle, Pierre. (1987) *Machine Translation at the TAUM Group*, in Margaret King (ed.), **Machine Translation Today: The State of the Art**, Edinburgh University Press.

Isabelle, Pierre. (1993) *Machine-Aided Human Translation and the Paradigm Shift*, in the **Proceedings of the Fourth Machine Translation Summit**, Kobe, Japan, pp. 177-179.

Justeson, J. and S. Katz. (1993) *Technical Terminology: some linguistic properties and an algorithm for identification in text*, Technical Report #RC 18906 (82591), IBM T.J. Watson Research Centre, Yorktown Heights, NY, 13 p.

Kay, Martin. (1973) *The MIND System*, in R. Rustin (ed.) **Natural Language Processing**, Algorithmics Press, New York.

Kay, Martin. (1997) *The Proper Place of Men and Machines in Language Translation*, in **Machine Translation**, vol. 23, pp.3-23.

Knight, Kevin. (1999) *A Statistical MT Workbook*, available online at http://www.isi.edu/natural-language/mt/wrbk.rtf .

Kuhn, Thomas. (1962) **The Structure of Scientific Revolutions**, University of Chicago Press.

Langlais, P., G. Foster, Guy Lapalme. (2000) *Unit Completion for a Computer-aided Translation Typing System*, in the **Proceedings of the Sixth Applied Natural Language Processing Conference** (ANLP-2000), Seattle, Washington, pp.135-141.

Macklovitch, E. (1986) *MT Trial and Errors*, in the **Proceedings of the International Conference on Machine and Machine-Aided Translation**, Aston University, Birmingham U.K.

Macklovitch, E. and G. Russell. (2000) *What's been forgotten in Translation Memory*, in the **Proceedings of the** Fourth **Conference of the Association for Machine Translation in the Americas** (AMTA-2000), Cuernavaca, Mexico, pp.137-146.

Macklovitch, E., M. Simard, P. Langlais. (2000) *TransSearch: A Free Translation Memory on the World Wide Web*, in the **Proceedings of the Second International Conference On Language Resources and Evaluation** (LREC-2000), Athens, Greece, pp.1201-1208.

Melamed, I.D. (1998) *Empirical Methods for MT Lexicon Development*, in the **Proceedings of the Third Conference of the Association for Machine Translation in the Americas** (AMTA-98), Langhorne, PA, pp.18-30.

Nagao, M. (1984) *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle*, in A. Elithorn and R. Banerji (eds.) **Artificial and Human Intelligence**, North-Holland, pp.173-180.

Nie, J-Y., M. Simard, P. Isabelle, R. Durand. (1999) *Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web*, in the **Proceedings of SIGIR '99,** Berkeley, CA.

Och, F-J., C. Tillman, H. Ney. (1999) *Improved Alignment Models for Statistical Machine Translation*, in the **Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora**.

Simard, Michel. (1998) *Automatic Insertion of Accents in French Texts*, in the **Proceedings of EMNLP-3**, Granada, Spain.

Somers, Harold. (1999) *Review Article: Example-based Machine Translation*, in **Machine Translation** vol. 14.2, pp.113-157.

Véronis, Jean and Philippe Langlais. (2000) *Evaluation of Parallel Text Alignment Systems – The ARCADE Project*, in J. Véronis (ed.), **Parallel Text Processing**, Kluwer, Dordrecht.

Appendix

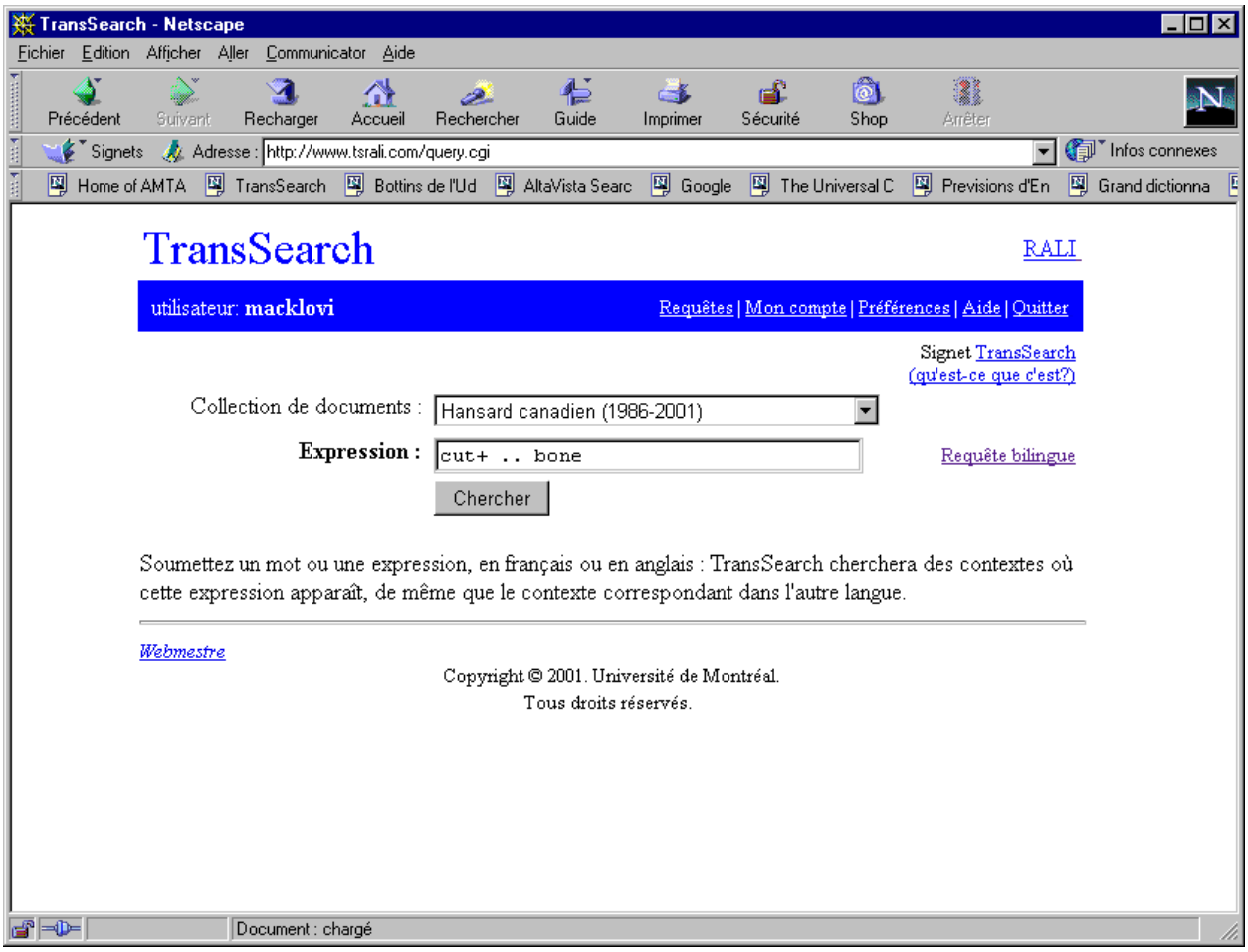

Figure 1: *TSrali.com* – the Web-based interface to the new version of *TransSearch*

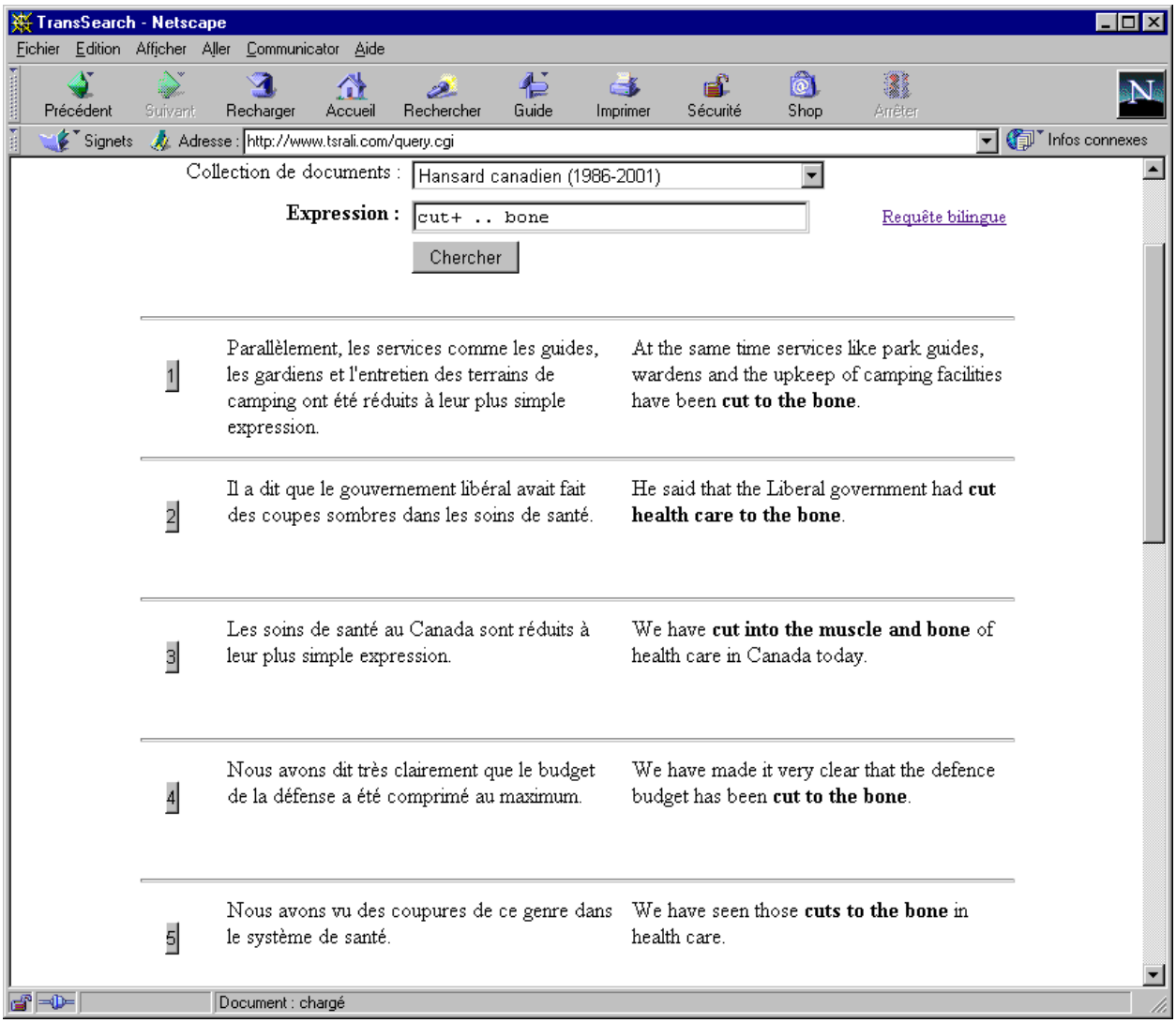

Figure 2: *TransSearch* – partial results of the previous query

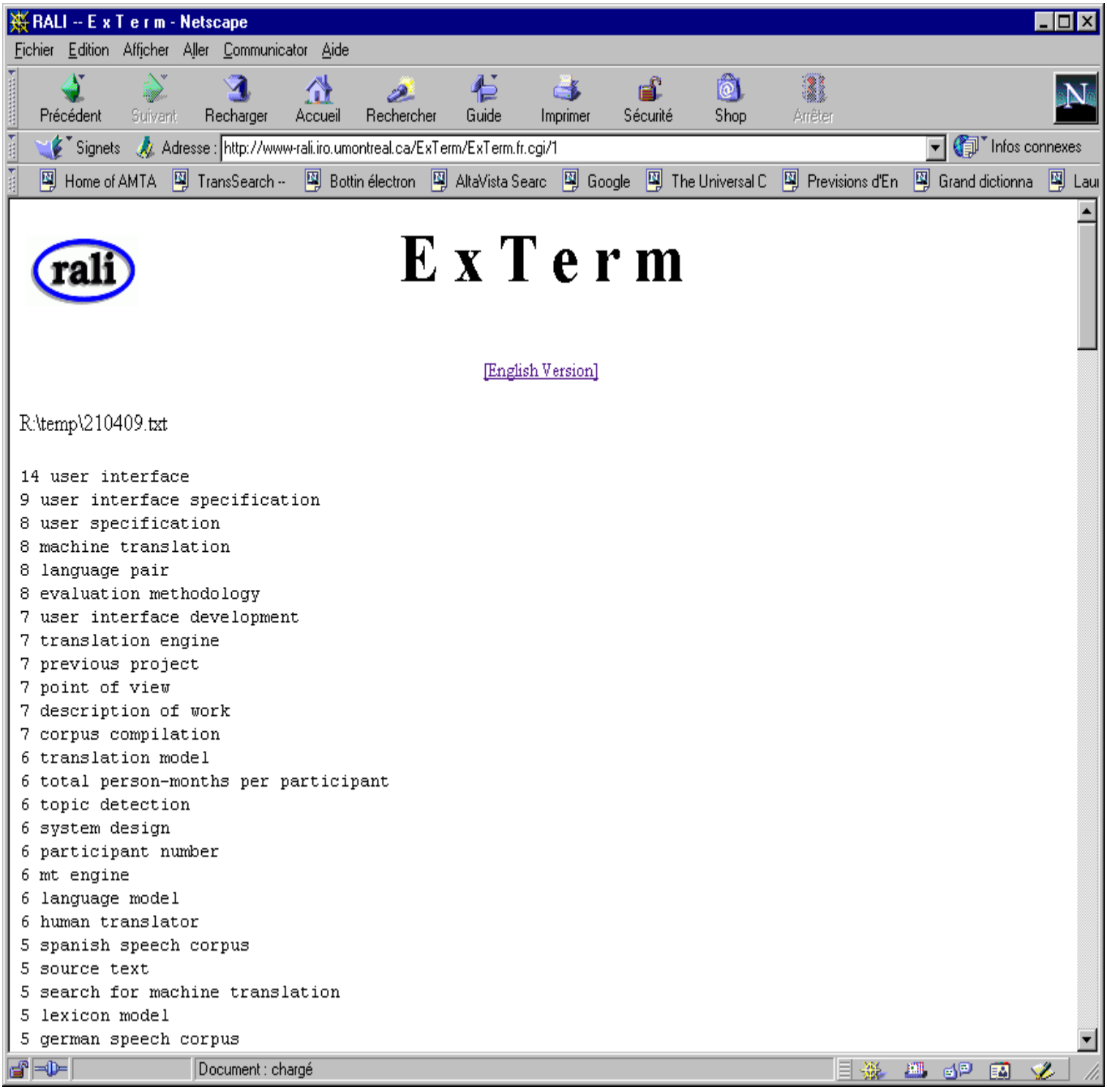

Figure 3: *ExTerm* – an example of monolingual term extraction

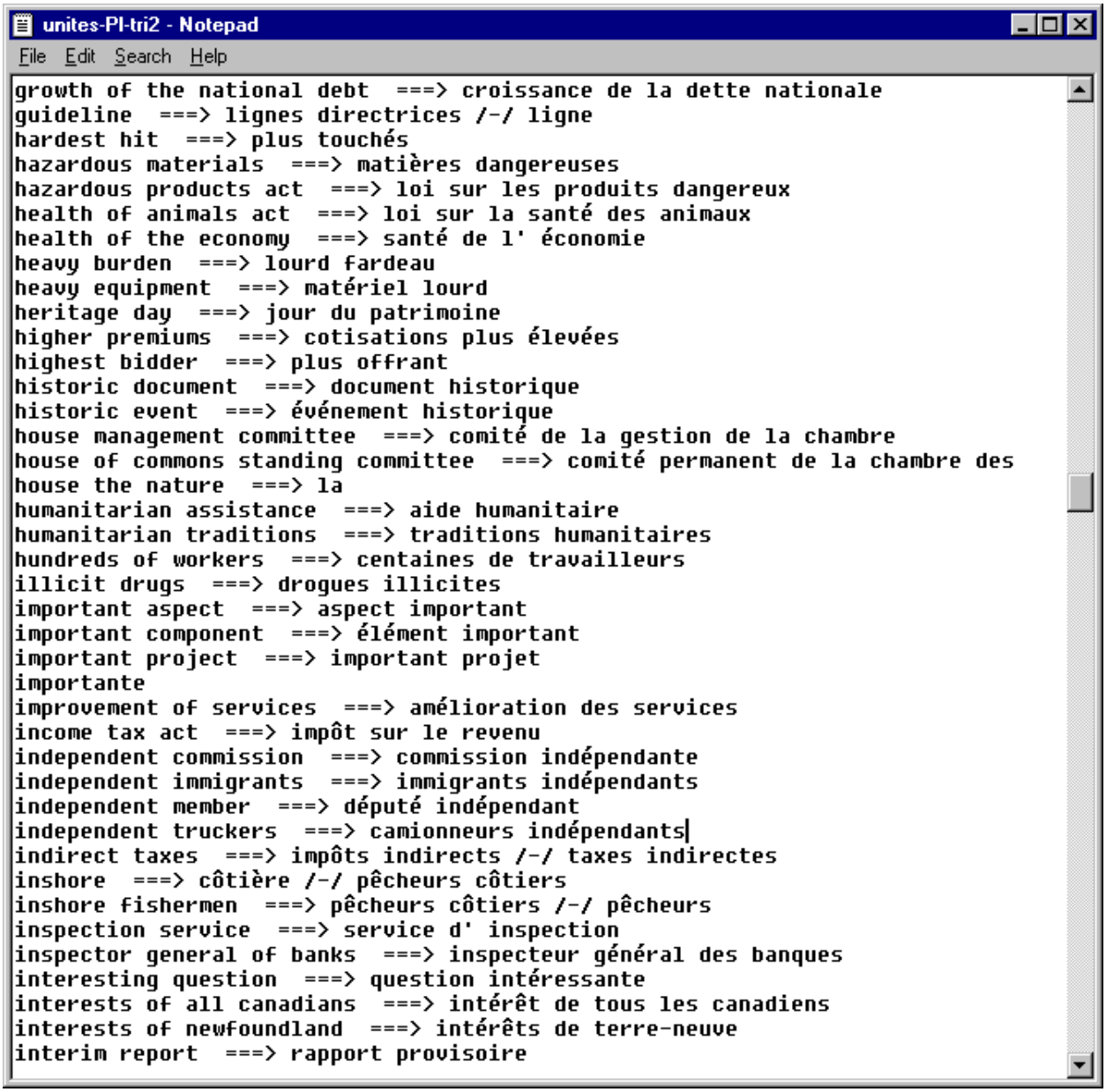

Figure 4: part of a bilingual lexicon automatically extracted from a parallel corpus

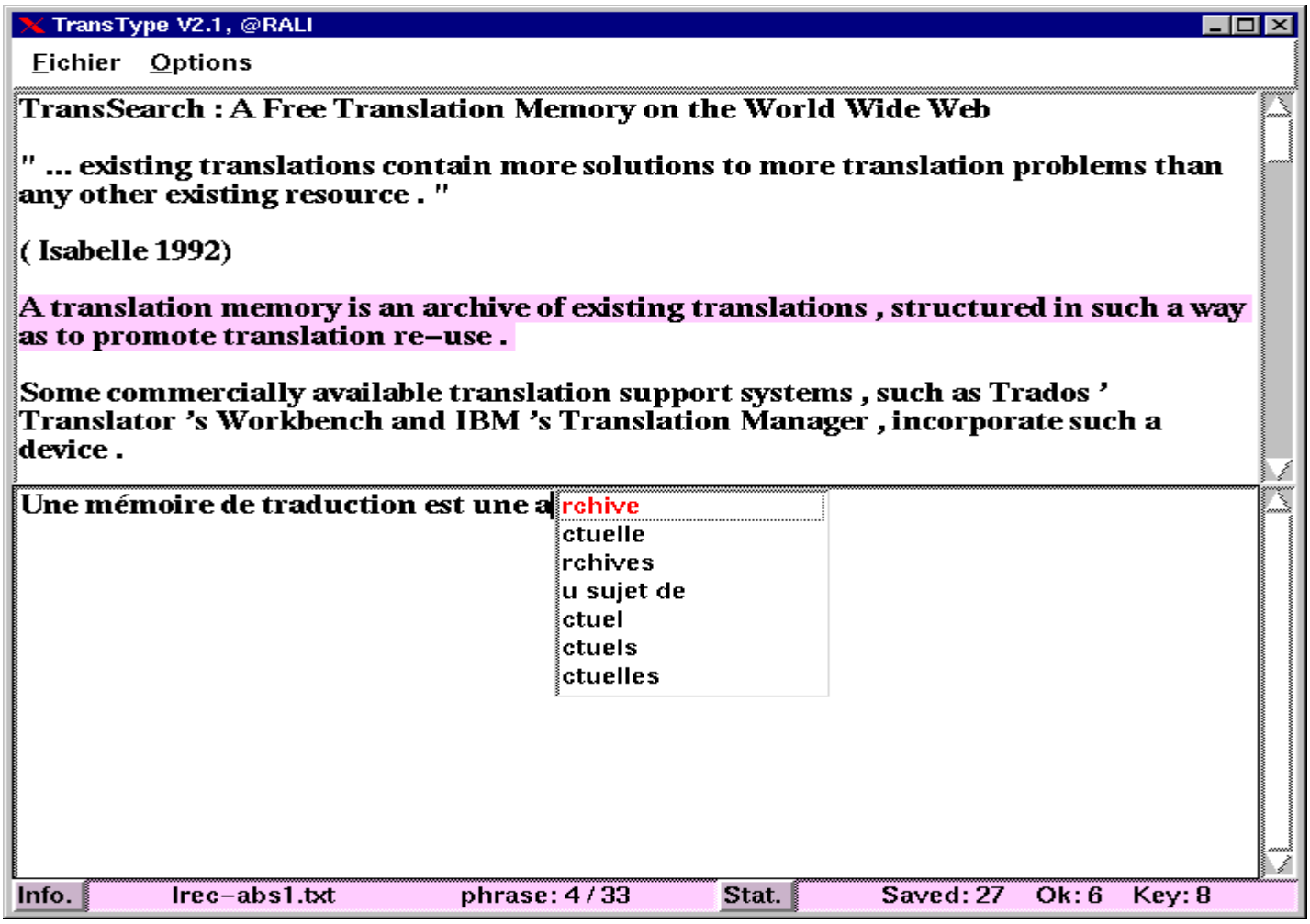

Figure 5: *TransType* – a snapshot of the current prototype. The system's suggested completions appear in the pop-up window in the bottom pane.