Bright Side of a Dark Age: Developments in Machine Translation, 1966-1992

Other than the fact that both are words, there doesn't seem to be much of a relationship between the word *hear* and the word *bravo*. One might conjure up various contexts to nestle them together in sentences, but as lexical items, in their most objectified states as word strings with dictionary definitions attached, they seem unrelated. They wouldn't sit anywhere near each other in an English dictionary, and the second term would probably find a home in dictionaries where the first would be absent. The same would hold for word lists and indexes organized according to other-than-alphabetic principles. But this is just brains and books talking. When computers are turned to the word association task, a relationship *is* discovered. The authors of "A Statistical Approach to Machine Translation," a 1990 watershed paper that instigated a revolution in the field of computer translation of languages, described the relationship as follows:

> As we would expect, various forms of the French word *entendre* appear as possible translations, but the most probable translation is the French word *bravo*.[1] When we look at the fertilities here, we see that the probability is about equally divided between fertility 0 and fertility 1. The reason for this is that the English speaking members of parliament express their approval by shouting *Hear, hear!*, while the French speaking ones say *Bravo!*. The translation model has learned that usually two *hears* produce one *bravo* by having one of them produce the *bravo* and the other produce nothing.[2]

The translation model and fertilities referred to here were products of Candide, a machine translation (MT) system developed at IBM in the late 1980s to translate from English to French. Instead of treating translation as a matter of interpretation or meaning transfer, IBM attacked translation as a statistical optimization problem.[3]

When Warren Weaver had first proposed using computers and statistical methods to translate in 1949, he'd conceptualized translation as a decoding exercise — English messages secreted in the wrappings of "strange" Chinese or Russian symbols. Unlike Weaver, who was curious about translation's broader contexts and practices, the IBM research group had no special interest in multilingual publishing or translation strategies. Rather, it was fresh off breakaway success in developing automatic speech recognition technology, and suspected that probabilistic methods, which had worked so well predicting speech strings, might do the same for translation. It was reconceived as English sentences that had been sent through a noisy communications channel that had garbled them into French. The basic premise was simply that "the characteristics of this channel can be determined experimentally, and expressed mathematically."[4] The approach was claimed to be "exactly analogous" to the speech recognition situation, in which the researcher has a sentence script and a time waveform and needs to establish a correspondence between them.[5]

The IBM paper gives a shout-out to Warren Weaver in its opening paragraphs. Peter F. Brown, sharing writing credit with seven other IBM research scientists, launched with a historical pronouncement that rooted the team's statistical discoveries in 1950s information theory and

placed them in lockstep with technological change. “Machine translation is almost as old as the modern digital computer,” Brown wrote, acknowledging that Weaver had proposed tackling it with statistical methods as early as 1949.[6] Researchers thereafter, Brown explained, had raised theoretical objections to statistical methods, and devoted themselves instead to rule-based approaches. Projects were thwarted even more and doubly, however, by the “impotence” of computers and the lack of machine-readable text.

By 1988, IBM was able to rush in with a suitably arousing linguistic matchmaking game that exploited the computer’s more contemporary endowments and uprooted the translational status quo with translation models described as “prodigal, spraying probability all over the place, most of it on ill-formed French strings.”[7] IBM’s reproductive language is, well, consummate. In today’s meme-speak: all your fertility are belong to us. The key to this particular romp — its map, and its territory — is a parallel corpus, a mass of aligned bilingual, human-translated text that is used to build a probability model of the translation process.[8]

A statistical approach to translation might seem odd, especially since this method has its foundations in signal processing. At the same time, theories of translation abound. Bilingual dictionaries, it can be said, theorize translation in terms of word and phrase equivalents. Their sets of ordered lists act as documentary arguments in support of the notion that words in one language align with, if not generate, words in another. If translation was a matter of word matching, translation would be a breeze for people and computers — just find and replace. IBM launched their investigation with this basic premise, then grappled mathematically with the fact that paired English-French translations don’t contain the same number of words. Motivating a theory of partial alignment, therefore, Candide calculated the probability that any single word corresponded to zero, one or two words in its translated pair. Because Candide was interested in moving from English towards French, IBM identified alignment as “the origin in the English sentence of each of the words in the French sentence.” As for the fertilities discussed in the excerpt above, they corresponded to “the number of French words that an English word produces in a given alignment.”

Once we understand the place of fertility calculations in channel modelling, the relationship between *Bravo* and hear seems far less unusual. “Hear, hear” and *Bravo* seem like a probable match — and without even venturing into French. Still striking, however, is the near even split between *entendre* and *Bravo*. In what cheery world of discourse would French *Bravo* appear so frequently?

The excerpt above shows IBM summoning the corpus to account for the anomaly — or, if one prefers, the regularity. Members of parliament (MPs) are mentioned because the corpus analyzed consisted of thirteen years (1973 to 1986) of Canadian parliamentary Hansard, the transcribed, edited and translated debates and proceedings from Canada’s House of Commons. IBM acquired the machine-readable corpus — containing three million sentences, a total of 100 million words — on magnetic tape sometime in 1987.

Whatever the strength of IBM's calculations around this particular example, their explanation gets one detail wrong. Their Hansard data, strictly speaking, says less about how English and French members of parliament spoke, and more about how parliamentary transcribers and translators wrote things down. *Hear, hear* is the perfect example, it being a shorthand, textual stand-in for any and all assenting statements (huzzahs and hurrahs surely among them) and sounds (including cheers, whistles, murmurs, claps and desk-poundings) that signalled agreement on the floor of the House. *Bravo!* does the same, but for Hansard's French version. IBM imagined it had transformed the speech of individual orators into variables for mathematical dissection. However, any uniqueness of voice had already been erased when House reporters turned the flow of parliamentary discourse into a document — through processing. "Hear, hear!" stood in for the redundant endorsements of unnamed, approving MPs. Such utterances were always attributed to "Some Hon. Members" (Some Honourable Members) in the final transcripts, before being shuttled down the line to the Canadian government's Translation Service.

The Bureau and Big Language

In other words, before Hansard was computed, it was published. The linguistically doubled deliberations of the state's legislative body were transcribed in shorthand, typed in longhand, translated verbally or in handwriting, typed again, printed and distributed. A closer look at the Hansard production chain, particularly as it was updated over the years, reveals a concise history of communication technologies. That the operational workflow delivered translations with such regularity and quality to readers under this regime allows us to entertain a metaphorical view of the Translation Bureau and cooperating government departments as a highly developed processing machine. In a no-fi, analog and bureaucratic way, one might say bilingual Hansard publication delivered on the promise of "FAHQT," or Fully Automated, High Quality Translation, the fantastical objective of 1950s MT research. That IBM trained its algorithm, dubbed the "Foundational Equation of Machine Translation," on this corpus, allows us to consider this an equally Foundational Database. Supporting this designation is at least one researcher, who referred at that time, if somewhat mockingly, to "their Hansard text" as "the Rosetta Stone, one might say, of modern MT."[9]

While most of the processes in the Translation Bureau workflow — printing, copying, shorthand, typing and dictation — have been dealt with in media studies — the translation step has been politely stepped over. One suspects that it deals too closely with meaning, or too straightforwardly with it to be of much interest. Generally speaking, materialist media studies in North America have assumed a "post-hermeneutic" position that looks at communication technologies in terms of transmission, processing and storage of data. Language, in strict versions of this framework, is looked upon as "a domain recalcitrant to internalization."[10] Machinations taking place inside translators' blackboxed minds, however routine and standardized they may be, do not easily enter such analyses.

A similar neglect is noted by translation scholars, but it claims its "ancillary condition" is precisely due to its perceived mechanicity, the fact that it is *not* creative. Still others claim that translation practice sustains its own invisibility by adhering to translational norms such as fluidity, accuracy and correctness, catalyzing on the linguistic level the kinds of media-representational illusions of transparency and immediacy that media scholars have worked hard to bring into view.[11] Nagging these positions is a vague moral or ethical stance around readability and responsibility; translation is a helpful stop-gap, but the "right" thing to do would be to learn that language yourself! Translation labour easily accompanies other neglected and often feminized care-taking roles that have undergirded the operation and upkeep of communication and information technologies.[12] Language labour might, too, though "acquiring" fluency is always described in the starkest capitalist terms, and championed in the most individualist. Seeing translation as information processing promises — or threatens — to unsettle assumptions about language's naturalness, about pathways to bilingualism and cultural expertise. Exacerbating the disregard for translation still further is the aura of suspicion and leakage that hovers over translation practice. If human translation is already neglected, secondary, uncreative, traitorous and possibly immoral, what chance does truly "mechanical" translation have?

It's a good time to reconsider translation operations, computational and non-, as constitutive of and embedded in informational flows. Translation's contexts and varieties are multiplying, translation is becoming entwined more resolutely within document production and circulation processes, and its supposedly "brainbound," fluently bilingual processes are morphing under pressures from crowdsourcing, outsourcing, and freelancing in ways that data mining and digital networks invent, sustain and make increasingly visible. We can take "meaning" for granted here, just as I know you can understand this sentence I'm writing (or pass it through Google Translate). A larger, and more interesting question, is how the meaning of translation, of information transfer via translation, and the organization of translation activity, is reordered in response to new pressures, through new technical arrangements and availability of tools. One part of this is the legacy of "older" linguistic media and the role they play in ongoing conceptions of translation (e.g. dictionaries, thesauri, schoolbook grammar, etc.) One way of recognizing these reorderings (which are both changes and continuities) is to look more closely at plans and designs to make translation more efficient, speedier, more accessible and less costly.

This chapter starts to chart a media archaeology of natural language processing by telling a stacking-doll story about machine translation. It traces the history and implications of the embedding of one kind of translation machine — one that's peopled, penned, printed and thoroughly political — into one that's automatic and algorithmic. Swallowed up inside statistical MT and radically decontextualized is the labour of the Canadian Translation Bureau. A layer deeper still is the tale of the Bureau's *own* pursuit to develop mechanized means to expedite the printing and delivery of Hansard. In other words, statistical MT embeds a media history. Each layer its own experiment in language engineering.

The notion of the "bigness" of language data (whether presumed, actual or mythologized), relative to the media formats, genres and contexts that contain and produce it, strains against this

story at every layer. A kind of format war starts to play out between "big" platforms (like English and French) and the means by which to automatically wrangle them in service of "small" documentary genres like debates transcriptions and weather reports. Fuzzy conceptual spaces like "general language" find institutional counterparts in "general services" departments, intensifying the problem and delineating the new uncomputable. A question worth exploring in this datalogical moment[13] is thus the fraught and codependent relationship between the design and operation of language processing tools and the particularities of the databases upon which they are conceived, and from which they draw. MT has always struggled with "Big Language" and its generic constraints. My hunch is that the history of MT may offer a surprisingly intelligible vantage point from which to grapple with some of the stakes of Big Data and its claims. At the very least, it offers a design and development story about language at scale, and the small scale actions that go, paradoxically, towards making it so.

The Quiet Era of MT Research

The short story told about early MT is that Warren Weaver of the Rockefeller Foundation got enthused about computers and cryptography and helped direct funding to the upstart field. Research flared brightly, but failed mightily in its attempt to get computers to translate Soviet science as fast as possible, and took an enormous amount of U.S. military funding along with it. While a few crude translation systems had been successfully installed at US military installations, mounting doubt led the NSF to request in 1964 that a committee be struck to assess MT progress, and to make recommendations to MT funding organizations. The Automatic Language Processing Advisory Committee (ALPAC), led by John R. Pierce of Bell Labs spent two years reviewing the nation's translation needs and costs, translator availability, human and machine capabilities and post-editing costs. MT came to an abrupt halt when the Committee's report was issued in 1966, making the unequivocal conclusion that "there is no immediate or predictable prospect of useful machine translation."[14] While Chairman Pierce did advise that research be continued in computational linguistics — "as distinct from automatic language translation" — the ALPAC Report brought an end to an era of MT.[15]

Developments in Canada complicate this familiar narrative in several ways: first, in its assumption that MT research ceased in 1966, second, in the conviction that MT research inherently derives from wartime cryptography and is bound up with military intelligence goals, and third, in the belief that it is a uniquely American and Anglophone story. MT historian John Hutchins has offered important counter-narratives, and newer histories of developments in the Soviet Union and France have helped fill the void.[16]

It's worth noting for the moment that scale was a crucial measure of achievement for MT, and figured prominently in the ALPAC decision to redirect funding from MT to computational linguistics. (No longer equals, MT falls under the latter field today.) Specifically, Pierce recommended that work at large scale continue, "since small scale experiments and work with miniature models of language have proved seriously deceptive in the past, and one can come to

grips with real problems only above a certain scale of grammar size, dictionary size, and available corpus."[17]

While Pierce was gathering his recommendations to cut support to MT, Canada was just revving its engines. 1965 is said to mark its inaugural MT moment,[18] and signalled the onset of the so-called "quiet" era of MT research.[19] This was the year the Canadian National Research Council (NRC) extended monies to MT research units at the universities of Montreal and Saskatchewan, and to the Cambridge Language Research Unit (CLRU) in the UK, marking an exception to that agency's standard policy of funding basic, rather than applied, research. The value of this endeavour, wrote F.T. Rosser of the NRC, is that "the results, if the project is a success, are of value and interest principally to other government departments, in particular to the Queen's Printer and the Translations Bureau."[20]

In fact, the project had already been under consideration for several years. A.F.R. Brown of the Georgetown University MT project had visited the NRC as early as November 1962, and shared preliminary work on a French-to-English system he had programmed in 1960. A write-up on the demo by a Translation Bureau employee is dismissive, and claims that Brown described it as an "expensive toy."[21] It's unclear who had invited Brown, but the evidence of strongest initiative after that point was from C.B. Watt, General Manager of the Canadian Government Printing Bureau.

Watt had been drawn to the idea after seeing a computer produce a hard-copy text from the tape output of a stenotype machine. While translation undoubtedly differed from stenotype, which followed a strict coding system, Watt envisioned a procedure in which typewriters also produced tape for feeding into automatic typesetters, which could then be fed into a computer that would translate it, produce a hard-copy and supply another tape in the second language for typesetting.[22] The hope was for end-to-end automation, such that "the whole operation of typesetting work, such as Hansard, as well as the translation into both English and French versions, could be done automatically."[23]

As it was, House of Commons debates were transcribed in shorthand by tag-teams of French and English reporters, who swapped each other out at 10-minute intervals. After every of these 10-minute "takes," each reporter buddied up with an amanuensis, who typed their French or English notes from dictation. A messenger picked up the typed copy and shuttled it to another building to another team for translation — French copy into English, and English copy into French. Audio kept flowing in ten-minute torrents, in mixed languages, the shorthand and typescript copy piled up, and the messenger made their rounds. Translated copy was proofread, then ferried on to the printer, where an editor revised and compiled, aligning English pieces with English and French pieces with French. Pagination was matched, and the two versions were printed.

Because Hansard was (and still is) not properly Hansard unless it exists in French and English versions,[24] the central function of this peculiar translation machine is the translation function, which is carried out by the personnel of the Canadian Translation Bureau, a centralized

translation service established in 1934 to manage all translation operations for the Canadian Government. Because Hansard is a working document, the commensurability of French and English versions is measured in timely printing and distribution. One of the "rules" that governs its production was, and still is, that the whole process be available for distribution by morning. The pressure on the Printing Bureau, the last in the chain, was acute, and may account for some of Watt's tenacity:

> What other parliamentary body […] produces overnight a separate Hansard of the full spoken and translated texts in more than one language simultaneously without a cut-off time for the Printer even if the House sits till dawn? … and also produces within 24 to 60 hours of a sitting a bilingual "Hansard" for each of up to 18 meetings of committees in a single day?[25]

Watt and the NRC asked the Treasury Board for approval, and a budget, for a feasibility study. The Treasury Board had reservations and internal memoranda advised against contributing substantial funds. Preliminary research had already been done in the U.S. and there was no sense in repeating these trials — all the more so as cost savings were not yet in evidence. Furthermore, the kinds of translations needed by Canadian parliament were not the "highly scientific works" that were (presumably) more suited to computer processing. Rather, Canadian parliamentary translation requirements consisted "of "literary" prose rather than technical work." The technology was suited "to more technical terms than it does to the wide variety of idioms that are to be found in intellectual and emotional types of writings." Furthermore, it had been pointed out to Watt that equipment was already available that could be used by Hansard staff to provide copy and tape simultaneously, and used for typesetting. Shouldn't this immediate step be taken, in advance of the proposed research? Watt was apparently "unable to say" why it could not. Remarkably, approval was granted, with notes to both "Protect the results of the study" and assure the Treasury Board that "findings will be Property of Gov't of Canada."[26]

A meeting was hosted at the NRC in July 1963, to discuss prospects for MT and convene a Canadian Government Machine Translation Planning Committee. Present at the meeting were Léon Dostert, Director of the Georgetown MT project, four representatives from IBM, NRC President Dr. Ballard, C.B. Watt, and Henriot Mayer, Assistant Superintendent of the Translation Bureau. Dostert and the Georgetown group were veritable rockstars of American MT, known for having put on the first public MT demo in 1954 — the "Kitty Hawk of electronic translation" — a showcase, too for IBM equipment.[27] The message was the same: advancements in printing and automation, combined with MT, might reduce the time lag in translations (especially those in French) and result in "more simultaneous release." A planning committee would be formed to work towards an "operational program in English-French machine translation in specified fields with a time objective." "If we can get this project off the ground," Watt confided to Mayer, "it will have world-wide implications and will bestow a great deal of credit on your department." Treasury Board approval was sought once again, and received in 1965, for a five-year research project with an estimated budget of $1.6 million.

1965 was also the year the Royal Commission on Biculturalism and Bilingualism — convened two years earlier to study and make recommendations in support of a more equal partnership between Canada's English and French populations — declared that the nation was passing through its greatest crisis in its history.

The daily marathon of Hansard production was just one of the Bureau's myriad translation tasks. It, too, sought a "general" solution, due to the encyclopedic nature of its mandated activity and problems acquiring and retaining translators. Those they could retain were either not sufficiently skilled in writing, or in the subject matter to be translated. The documents needing translation were of every kind and genre, comprising speeches, fact sheets, statistics, correspondence, legislation, minutes, clippings, press documents, "love letters" from citizenship applicants, as well as reports, articles, proposals, bulletins, descriptions of machines and equipment, tenders, contracts and maps from government ministries as diverse as Fisheries and Forestry, Mining, Agriculture, Natural Resources, Trade, Defense, Citizenship and Immigration, Public Works, etc. At the time, English remained the language of conception and composition throughout the public service. It was translation, which comprised techniques, practitioners and documents, that produced linguistic parity.

The delivery of French versions of this range of documents consistently lagged, and the Translation Bureau set to rectifying the situation, with increasing pressure from Francophone media and from departments eager to correct the disparity between French and English, at least as measured in page counts. A new "ideal," modeled on Hansard production, increasingly asserted itself in support of French-English alignment: a work would not be considered "published" until English *and* French versions appeared. "Bilinguisme simultané," in the form of same-time publication, would strip French of offensive second tier status. This emphasis on linguistic synchronization intensified again when a mechanism for simultaneous translation of verbal discourse was introduced to the House of Commons in 1959. Along with wiring, boxes, switches, earpieces and elaborately carved booths, the installation of the so-termed "simultaneous translation system" siphoned off the best House of Commons translators for training in interpreting. While it succeeded in injecting more French into debates, it ignited controversy over the fate of second language learning given the lazy appeal of switch-activated language. Once adopted, very little lexical distinction was made between text translation and speech interpreting; "simultaneous translation" in any of its governmental manifestations was thus easily conceived as a two-channel, multimodal language duplication service whose product was bilingualism. As Prime Minister Diefenbaker explained, and also in 1965: "My contention was that parliament is and must remain bilingual and the only way to make it effective is through the instrumentality of a system such as that."[28]

One Single Paragraph

Two of the research teams awarded contracts in NRC's new venture already had long pedigrees in MT research. One contract was with Andrew D. Booth and Kathleen H.V. Booth, a crystallographer and mathematician recently relocated from Birkbeck College to the University

of Saskatchewan. The other was with Margaret Masterman, director and founder of the Cambridge Language Research Unit, one of the earliest MT research centres.

Masterman's Pilot Research Proposal submitted to the NRC states its objective in its title: "for the comparative analysis and processing of modern English and French Canadian official documents; this analysis being made with the further purpose of establishing a mechanical translation procedure between them." The proposal is rich and detailed, and describes the bulk of Masterman's and the CLRU's major discoveries from the previous decade, enumerates their pros and cons, traces their turning points and explains how they've contributed towards new insights. The discoveries they were then combining into a semantically-focused translation technique included the "Mechanical Pidgin Translation" method, Masterman's influential thesaurus approach, which codes input text to thesaurus heads and translates semantically, i.e. "non-literally," multilingual information retrieval, and a means of identifying salient pieces of discourse. The CLRU approach did not accept "garbage" in its systems — that is, it would not produce something less useful to the end-reader than the original version that s/he was initially unable to read — nor did it tolerate building systems whose output required post-editing. Masterman's plan of attack for mechanical translation from "official-government-document" English into Canadian French would actualize the next phase of CLRU's experimentation. Proposed was the following:

> the construction and testing of a computer-program to detect the messages of A. ONE SINGLE PARAGRAPH of Official Canadian-Governmental English, and B. ONE SINGLE PARAGRAPH of corresponding Canadian French. (B. being the Canadian Government's official translation of A.), in order subsequently to compare, correlate, and inter-transform the messages.

Additional dictionary compiling, theoretical and practical work, would certainly be involved, but the team would limit itself to delivering the single paragraph as outlined. Masterman estimated this would require five team members for a two-year period. The CLRU contract lasted for those two years with the NRC, but was not renewed. The NRC forwarded the proposal to Mayer of the Translation Bureau in 1965, stressing that though the work was "of a fundamental nature," it was sure to stimulate the project, as the group was very good "in producing new ideas."[29] For all the inventiveness and experience in this outline, the single paragraph promise might have seemed absurd to Mayer, crushed regularly by thousand-page demands.

A 1967 paper by Masterman (still supported by the NRC) may have been even further off-putting given the political climate. The system proposed a man-machine interaction system that helped an "Englishman at a console but who does not know any French" produce idiomatic texts. By engaging in question and answer sessions with the machine to develop a picture of a sentence's underlying semantics, the operator effectively pre-processed "English input into Frenchified shape." Such a procedure aimed to produce a translation that "accounts for the actual non-literal translation which was actually made by the official Canadian Government Translator." Most

translators in the government service were Francophones, and a one-way simulation designed to assist only Anglophones was politically misguided and threatened French job security.[30]

While the Booths' contract lasted several years longer, they claim that political differences and biases caused its termination.[31] They came to the University of Saskatchewan in 1962 and became involved with C.B. Watt and the Hansard translation endeavour soon after arrival. They had both been involved in different aspects of computer engineering and design, programming and MT for nearly two decades at Birkbeck College, London. Unlike Masterman, their approach assumed that translation output would require post-editing. The team produced a program and demonstrated it at the NRC in 1971.[32]

A paper on first steps undertaken by Kathleen, containing 15 pages of statistical tables showing basic frequency deviations between French and English,[33] was included in *Machine Translation* (1967), a volume edited by Andrew that brought together work done beyond the US — projects untouched by ALPAC's "great gloom." Booth claimed in the introduction to be unaware of any other collection on MT that featured "any relevant statistical data about any language."[34] His hope was that other groups would be inspired to make similar data available.

The Saskatchewan project maintained that MT should involve attention to structure of words and structure of grammar, but that words should not be arranged as in a conventional dictionary, but according to frequency. Some grammatical tendencies were more likely than others, in terms of frequency of occurrence, and Booth's team had discovered ordering differences between English and French by analyzing sentences computationally. The program rested on a principle of word-for-word translation, after which it tried to handle an increasing number of complications, "the noun-adjective-adverb situation, the pronoun situation." The result was a program called MT6, "quite a potent program."[35]

Aiming for real results in finite time, the Saskatchewan team limited themselves to a particular text genre — "government" type documents — which they defined in opposition to literary composition and transcribed speech. They would seek "a good approximation," not perfection. The latter restriction meant their approach would rely on editors "to tidy up the machine output, select the most apposite translation of a word when more than one is available and, hopefully, correct cases where the program has gone astray and misinterpreted a sentence."[36] The program consisted of a Reader, Analyzer and Store, which handled input text, performed sentence and work segmentation and arranged the constituents in high-speed memory. The dictionary part of the program was "on-line," stored on direct-access disks or drums for high-speed searching, and was arranged for search and storage efficiency (e.g. English words were stored separately from their translated counterparts, which tended to be longer words). An isolation routine performed a search for idioms. Another phase involved statistical parsing, which assigned individual words one of sixteen grammatical codes and provided the French translation (with 4% error).

The program wasn't complete, "and indeed probably never will be if the criterion for completion is that no improvement is possible."[37] The mechanism worked on a laundry list of grammatical

transformations of the type one learns in grammar books: look for verbs and verb groups, assemble them where necessary; rearrange nouns and adjectives; "process" verbs by tagging their subjects, compound verbs are "associated," negatives are "rearranged" and verbs are "inflected" according to tense and subject. The arrangement of these steps was admittedly arbitrary, having been "dictated by the fact that the program "just grew" that way."[38] They were able to anticipate and exploit certain consistencies in the corpus to minimize semantic ambiguities. Their training corpus — 25,000 words of *Canada Year Book 1962* — was written entirely in the third person, and its lack of pronouns and passive constructions minimized problems.

The translation part was written in COBOL, which had been the only high-level language suitable when they began the project. The program contained 1500 statements, and the translation time for 1800 words was 5 minutes. As the aim of their work was "to provide assistance to the overworked Government translation services in Ottawa," the team wanted to

*NORTH *AMERICA COMPRISES SIX MAIN NATURAL REGIONS WHICH ARE BOTH PHYSIOGRAPHIC AND GEOLOGICAL BECAUSE THE AGES, KINDS AND STRUCTURES OF THE UNDERLYING ROCKS DETERMINE THE NATURES OF THE LAND SURFACES. *KNOWLEDGE OF THESE REGIONS IS IMPORTANT BECAUSE THEIR GEOLOGICAL CHARACTERISTICS HAVE MUCH INFLUENCE ON THE SUITABILITY OF DIFFERENT AREAS FOR SUCH ACTIVITIES AS AGRICULTURE, MINING, PETROLEUM PRODUCTION AND SPORTS, AND CONTRIBUTE AS WELL TO THE VARIED SCENERY OF THE COUNTRY. *THE SIX REGIONS ARE: THE *CANADIAN *SHIELD, A VAST AREA OF ANCIENT ROCKS THAT IS MAINLY IN *CANADA; THE *INTERIOR *PLAINS AND *LOWLANDS, THE LARGEST AREA OF WHICH EXTENDS THROUGHOUT THE MID-*CONTINENT FROM THE *GULF OF *MEXICO TO THE *ARCTIC *OCEAN; THE *APPALACHIAN *REGION, MAINLY IN THE *UNITED *STATES BUT ALSO FORMING AN IMPORTANT PART OF *EASTERN *CANADA; THE *CORDILLERAN *REGION, EXTENDING ALONG THE ENTIRE WEST COAST OF THE *CONTINENT; THE *ATLANTIC *COASTAL *PLAIN ALONG THE EASTERN SEABOARD OF THE *UNITED *STATES; AND THE *INNUITIAN *REGION, A MOUNTAINOUS BELT IN THE *ARCTIC *ARCHIPELAGO. *CANADA INCLUDES PARTS OF FOUR OF THESE REGIONS AND ALL OF THE *INNUITIAN *REGION, BUT NONE OF THE ATLANTIC *COASTAL *PLAIN.

L'*AMERIQUE DU NORD COMPREND SIX R1EGIONS PRINCIPALES NATURELLES/PROPRE/ QUI SONT 2A LA FOIS/ AUSSI BIEN QUE/ PHYSIOGRAPHIQUES ET GEOLOGIQUES PARCE QUE LES 3AGES, LES SORTES ET LES STRUCTURES DES ROCHES SOUS JACENTES/FONDAMENTAL/ D1ELIMITENT/ D1ECIDER/ LES TERRAINS/NATURE/ DES TERRES DE SURFACE. DE LA/LA/ CONNAISSANCE DE CES R1EGIONS EST IMPORTANTE PARCE QUE LEURS CARACT1ERISTIQUES G1EOLOGIQUES INFLUENT BEAUCOUP SUR L'ADAPTATION DE ZONES/SUPERFICIE/ DIVERSES POUR DES ACTIVIT1ES TELLE QUE DE L'/L'/ EXPLOITATION MINI2ERE, DE LA/LA/ PRODUCTION P1ETROLI2ERE/DU P1ETROLE/ ET DES/LES SPORTS, ET FOURNISSENT/CONTRIBUER/ EN PLUS AU PAYSAGE VARI1E DU PAYS/R1EGION/. LES SIX R1EGIONS SONT: LE *BOUCLIER *CANADIEN, UNE ZONE/SUPERFICIE/ VASTE DE ROCHES ANCIENNES QUI EST PRINCIPALEMENT/ DANS L'ENSEMBLE/AU *CANADA: LES PLAINES INT1ERIEURES ET LES BASSES TERRES, LA PLUS GRANDE ZONE/SUPERFICIE/ DE QUI S'1ETEND/PROLONGER/ 2A TRAVERS LE CENTRE DU CONTINENT DU/DEPUIS/GOLFE DU *MEXIQUE 2A L'OC1EAN ARCTIQUE: LA R1EGION *APPALACHIENNE, PRINCIPALEMENT/DANS L'ENSEMBLE/AUX *1ETATS *UNIS MAIS FORMANT/ ASSURER/ AUSSI UNE PARTIE/1ETENDUE/ IMPORTANTE D'*EST CANADIEN: LA R1EGION DE LA *CORDILL2ERE, S'1ETENDANT/PROLONGER/ LE LONG DE TOUTE LA C3OTE OUEST DU CONTINENT: LA *PLAINE C3OTI2ERE DE L'*ATLANTIQUE LE LONG DU LITTORAL EST/ORIENTAL/ DES *1ETATS *UNIS: ET LA R1EGION *INNUITIENNE, UNE ZONE/CHA3INE/ MONTAGNEUSE DANS/AU COURS DE/EN/ L'ARCHIPEL ARCTIQUE. LE *CANADA COMPREND DES/LES/ PARTIES/1ETENDUES/ DE QUATRE DE CES R1EGIONS ET TOUTE LA R1EGION *INNUITIENNE, MAIS NE-AUCUN DE LA *PLAINE C3OTI2ERE DE L'ATLANTIQUE.

Note: Numerals represent accents as follows:
1 acute
2 grave
3 circumflex
Slashes / / indicate alternative translations
* indicates capital letters.

Fig. 1

assess how understandable their results would be to someone without access to the original document and who did not necessarily know English. This interest in "understandability" was quickly passed over and reconceived as a correction exercise performed by well-educated French persons. The French reader became a bookish, red-penned grader, and an article ostensibly about "English-French Translation on a Computer" was thus only part about computer results, the other part a reading and comprehension test of French computer writing.

For the test, forty sentences from a section on Geology were taken from *Canada Year Book 1962*. Instead of using Quebeckers as reader-correctors (if well-educated, they probably knew English), the researchers chose subjects from France, most of whom had some university education and half of whom had had no instruction in English. Quantitative assessment of the readers' performance, it was acknowledged, was difficult and potentially misleading. As a measure of general performance, the readers understood 24 of 39 eligible sentences correctly, and over half grasped the meaning of 31 out of 39. "Four particular sentences were responsible for over half the errors made." At this point, one notices that the assessment started to venture away from being an analysis of system errors and results, and had started ascribing error to the human comprehender… and potential means for eliminating these comprehension gaffes.

The authors planned to write a follow-up paper that delved into the post-editing task results by categorizing editing errors as "highly probable," "possible" and "unanticipated" and calculating the percentage of post-editors "who actually went wrong." If translations rendered faithfully the content of the original, what exactly was it "there then in our output which caused some post-editors to make choices between alternatives and changes which caused the corrected output to do other than convey the meaning of the English text?" The computer output, from this new perspective, is a meaning prompt. If better constructed, the route from computer output to ideal "correction" might be easier to discern. The problem now was not so much how to improve the computer output, but how to improve the way the post-editor apprehended the results. The successful program had been stripped of stylistic requirements, but machine writing delivered "sense." What used to be called syntax (certainly resolvable by human translators) would perhaps start being described as machine "style," machine authoring that people could learn to read.

Montreal: Starting from Scratch

It was not only politically advisable to include a Quebec institution in the NRC initiative, but linguistically and culturally so; a project intended to translate between French and English should have French speakers and linguists in the mix. The NRC threw a third portion of funding support behind a linguistics division opened that same year at the University of Montreal, called CETADOL (*Centre de Traitement Automatisé des Données Linguistiques*), under the directorship of linguist and second language teaching specialist Guy Rondeau. While Rondeau didn't specialize in MT, he'd taken an interest, having visited MIT and translated MT researcher Victor Yngve's 1963 book, *An Introduction to COMIT Programming*, a string processing language for use by linguists, into French. Rondeau's linguistics connections, recruitment skills and fund-

raising abilities helped convince the NRC to put their funding and faith in the non-pedigreed Montreal group. Furthermore, Rondeau had completed his doctorate in France and made connections with linguists there whose influence and participation would be central to the research direction taken.

As Rondeau had his fingers in several computational linguistics pies, CETADOL focused early efforts on broad and basic research, tackling a range of natural language processing questions, including English morphology, syntax and grammatical classification. In other words, the team did not know how to proceed with the specific task at hand. As Harris put it,"We were really fishing around."[39] Rondeau had divided the team into three sections — English, French and Computer — but there was no communication between the sections. The NRC called a meeting at the end of the first funding year, bringing together members of all three teams. Harris had been analyzing a corpus of English phrasal verbs by hand and presented his findings. Yorick Wilks, attending the meeting on behalf of the CLRU, responded with "But could you now please tell us how you propose to put it on the computer?" Harris was "absolutely stymied." Harris found a programming course upon his return to Montreal with the help of computer centre director, and early code poet, Jean Baudot. It wasn't easy to find a suitable course as the bulk were oriented towards people with experience in science and mathematics. Harris joined a course for librarians and learned to program in Fortran. This was a problem for CETADOL as well, which ran into recruitment problems in its second year. There was a dearth of computer expertise, and "while linguists were not too difficult to find, we had great difficulty in obtaining programmers."[40]

In contrast to CLRU and the Booths, the inexperienced CETADOL team was mostly starting from scratch, and worked on exploring and defining its task in its earliest progress reports. Two papers, one by André Dugas, and one by Brian Harris, try to pin down research questions, select best methods and delimit their research tasks, but end up confronting political and philosophical questions.

Dugas' paper sets out to explain the "pivot language" method for translation, an approach CETADOL was exploring following their colleague Bernard Vauquois at the University of Grenoble. When one adopted a pivot language in MT, Dugas emphasized, one had in mind a device or apparatus which stored all the information generated from the analysis of natural language data, as well as information that *did not emerge* from the analysis of the original data, but which will serve in the "manufacture" of the target language. The pivot was an intermediate representation that sat between two natural language strings. Dugas felt the pivot-language approach was absolutely necessary for MT, but provided examples that suggested he was concerned about the way the method was programmed. Pivot language information had to be coded, after all; a too-crude MT system could impact negatively the "manufacture" of French — as it sat in the target language position. The translation of "Provincial governments" to "les gouvernements provinciaux" would seem to be easily accomplished with a word-for-word approach. A simple operation would add "the" at the sentence "manufacture" stage and swap around the word order. However, a second version — and the preferred translation — would be much harder to program. Quebec in 1967 was struggling for political independence from

Canada. For francophone readers to whom this translation would be addressed, "Provincial governments" should be rendered as "le gouvernement du Québec et les gouvernements des autres provinces."[41] A footnote reminds the reader that the project involves the study and translation of Hansard texts. Hansard translators would not enter a translation into parliamentary record that acknowledged Quebec's claim to sovereignty in this way, both grammatically and programmatically, by differentiating one government from all the others with a conjunction and by making it routine. For Dugas, the pivot language offered a way to intervene in and re-program politics. Rather than reproducing longstanding equivalents, new political formations could be concretized through translation, "manufactured" by updating outdated sentence-generational rules to reflect new political circumstances.

Brian Harris's paper tried to delimit the aims and scope of their task of developing a mechanical translation system. He identified MT as "a practical task with tangible results" and associated it with applied linguistics, not theoretical linguistics. Whereas "general linguistics" might try to seek a grammar that would serve as a theory of language applicable to all languages, "a grammar for use in machine translation should be chosen according to its fitness for that task." The result might be grammars that were radically different than those sought or proposed by theoretical linguists. Harris was pointing out that MT developers were more concerned about performing potentially narrow translation tasks than forwarding lofty propositions about the nature of human language (or sets of all possible sentences). Grappling with the size of human languages and wrangling some piece of it to work with was the key challenge. Despite the distinction just made, Harris could still imagine an *applied* linguistics that took the entirety of a language into account. Among the parameters to consider in developing MT recognition grammars, designers needed to consider "whether translation is to be from a narrow segment of the source language, or from a broad segment, or indeed from the whole spectrum of the source language." The imagined scenario was unlikely, as it would "probably require as many grammars and programs as there are speakers of the language." The narrowing down had to be done by medium or genre. Thus, "all projects begin with a declaration of intent, which may be to translate 'scientific communications' or 'newspapers'."

Even these domains were too broad, too unwieldy. (Once narrowed, did they scale?) "The more cautious," Harris suggested, "tackle much narrower segments." The project in Grenoble, for instance, focused on a subset of language, only on titles of scientific articles. They discovered that titles had a syntax of their own, a structure that differed from the rest of the scientific text. Any grammar that analyzed both would need to account for the particularities of each part. The smallness and particularity of a given grammar, Harris was careful to point out, did not diminish its usefulness. The point was to recognize different elements of language and the need to create different mechanisms to translate them. Such tools had value beyond translation contexts, too, for instance, a title-analyzer "would be immensely valuable in automatic documentation work." Harris' foreshadowing fused poetry and negentropy as he contemplated a generalization not around the chaos of language's myriad contexts, but on its narrowness: "There may even be some coefficient by which the power of any grammar increases in proportion to the narrowness of the segment of the language that it 'generates'." Harris's prophesy was an early warning about

the appetite for Big Language. “For those who would try to chew too much,” he writes, “automatic translation may be like the gateway to the Inferno…”

> *Per me si va nella città dolente …. (63)*

Harris trailed off this section with a line from Dante. Left unasked: what recognition grammar could be devised to account for *multilingual* text?

A question that needed urgent tackling was the size of the corpus necessary to deduce a grammar. Here, Harris’s earlier observations regarding narrowness of a grammar geared to a particular medium or genre of translation suddenly backfired, because “the more extensive a grammar needs to be […], the bigger the corpus has to be, too.” Samples could be collected from here and there, samples could be invented, or one could “take a specific text and limit the grammar to what that corpus proves.” Each method would generate differently sized grammars, but they would be equal in power and validity. Invented examples — thought tainted by researcher intuition, thus less disciplined — were no less valuable. Preconceptions were unavoidable with every method, Harris conceded, since “everybody comes to the grammatical analysis of a familiar language with a preconceived ‘model’ of the language and a wealth of ‘data’ already in his head which he cannot shut out of his deliberations.” The fact, while obvious, needed stating: you had to know a language to program its grammar. (65-66)

The problem of big language/small corpus was this: how could one generalize and be particular at the same time? Lexical and syntactic regularities were readily discoverable at the sentence level, but one needed thousands more sentences in order to happen upon “less common ones that turn up only once in ten, or ten thousand, sentences.” Harris’ conviction about the power of a narrowly defined grammar suddenly flipflopped: “if a corpus is small, the grammar derived from it is likely to analyse usefully only the most common mechanisms of the language, plus a virtually useless pot-pourri of the less common ones.” (66) One was seeking a perfect model for a system that seemed unsystematizable. Harris compared a sentence translated word-for-word from English to French, then considered how it could be improved after applying a few grammatical rules. Even without improvement, he concluded, “The sentence is now quite comprehensible in French.” “With a little practice using this type of translation, a French reader could even cope easily with the first version.

While such outputs tended to be called “intermediate representations,” stepping stones to a more final, truer, or less conspicuous kind of output, it was equally true that MT generated new kinds of language, again necessitating end-reader training in a new kind of reading. Translation failure — by design, but seemingly by computer — could be rewritten as success when it was offloaded to users who learned to read this new computer-generated language.

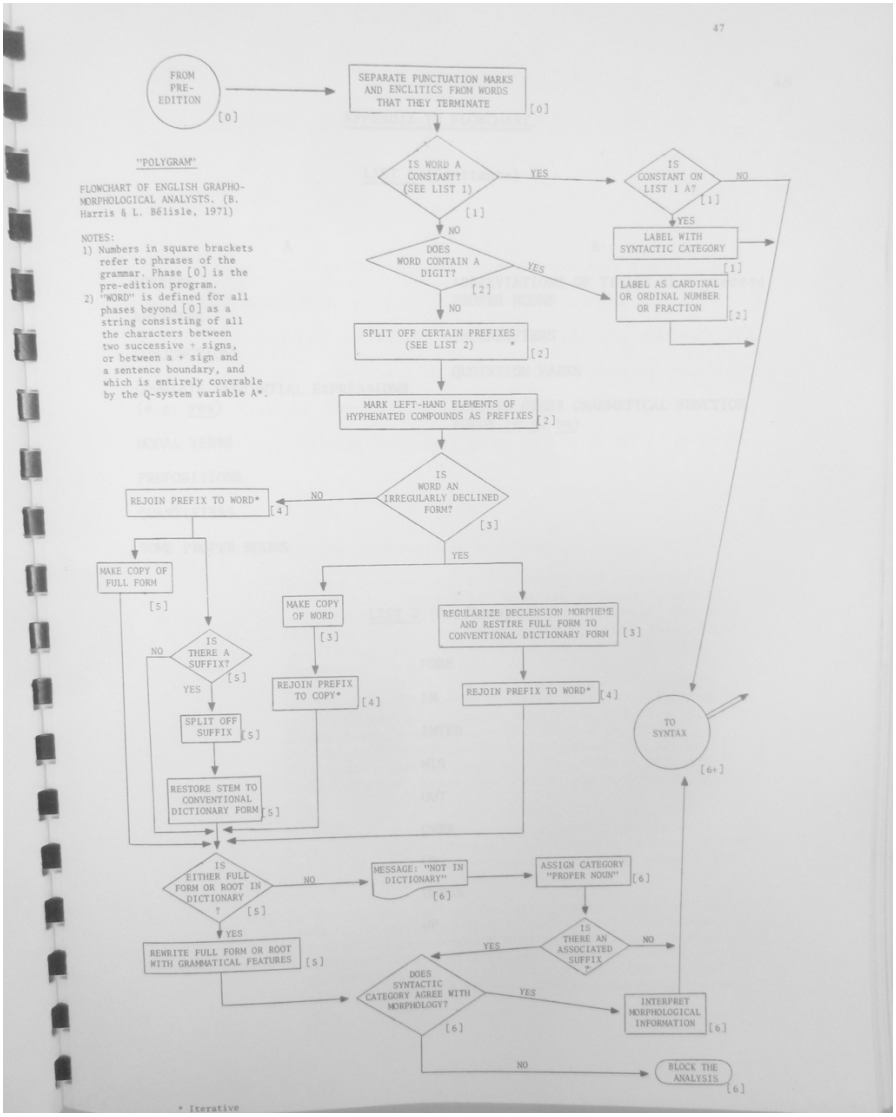
Subsequent quarterly reports revealed that CETADOL researchers were engaged in an enormous amount of general linguistics research in preparation for translation, but translation was still not underway. The situation changed dramatically when Rondeau left the University of Montreal in

1968 and was replaced by Alain Colmerauer, whom Rondeau had successfully recruited to the project from the Mathematics Department.[42] The first translations were performed in January 1970, "a limited, but vital experience," the first implementation of Colmerauer's so-called Q-systems, a programming environment created for the project.[43] Reflecting on Q-systems development, Colmerauer frames it not in terms of a tool in service of MT, but as "the result of a first gamble: to develop a very high-level programming language, even if the execution times it entailed might seem bewildering." A Q-system is defined in TAUM's 1971 report as "a set of rules which allows certain transformations to be carried out on oriented graphs." Colmerauer described the challenge, and Q-system's answer to it, as follows:

> It is difficult to use a computer to analyze a sentence. The main problem is combinatorial in nature: taken separately, each group of elements in the sentence can be combined in different ways with other groups to form new groups which can in turn be combined again and so on. Usually, there is only one correct way of grouping all the elements, but to discover it, all the possible groupings must be tried. To describe this multitude of groupings in an economical way, I use an oriented graph in which each arrow is labeled by a parenthesized expression representing a tree. A Q-system is nothing more than a set of rules allowing such a graph to be transformed into another graph. This information may correspond to an analysis, to a sentence synthesis or to a formal manipulation of this type.[44]

Q-systems finally gave the team (by this time, renamed TAUM — Traduction Automatique à l'Université de Montréal) a programming language they could work with. As Brian Harris recalled, "You didn't have to know anything about how the computer worked, you just had to use this to write programs." Another feature of Q-systems that lent itself to MT was that the output of one Q-system program could be submitted automatically through to another Q-system program. "We could do English grammar, pass it through the pivot, and pass it on to the French side." The system was the input for the next system. English sentences were run through 15 Q-systems on their way to becoming French (two for morphology, one for English analysis, two for transfers from English structures to French structures, one for synthesis of French, and nine for French morphology"[45]). Q-systems were also reversible, meaning that the same Q-system could describe, for instance, the analysis and synthesis of the same sentence.[46]

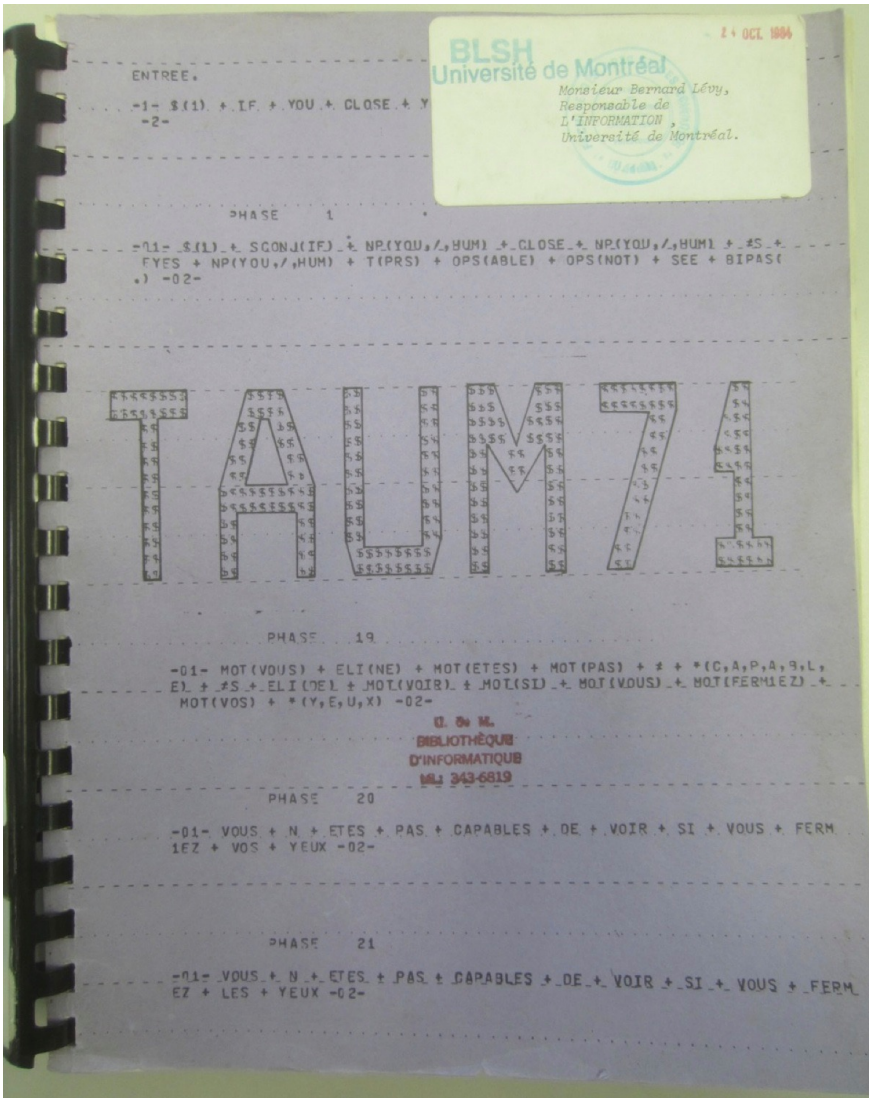
The Q-systems were "the star" of the annual meeting with the NRC and the Translation Bureau in 1971 although, in Harris's estimation, none of the bureaucrats understood anything about computational linguistics or MT. Although TAUM 71 was considered a prototype, the government agencies were getting nervous that TAUM still hadn't delivered any working systems. They finally brought a "computer expert" to assess that year's activity. Harris had been pursuing a sideline project, a primitive system that explored some notion of translation memory, which he called "The Transformulator." The expert told Harris he'd "saved the project … it worked!"



Flowchart prepared by Harris describing a grapho-morphological analysis of English suffixes prepared for Q-systems.

Colmerauer returned to France, leaving Q-systems with TAUM ("Q" stood for Quebec, after all). Richard "Dick" Kittredge, an American linguist, assumed the directorship, and the team published another major report in 1973. Q-systems had now been in operation for three years, and TAUM was continuing to exploit its power and efficacy, trying to work with translations of larger texts, with a focus on notoriously difficult longer phrases. It's worth nothing that Kittredge's "Introduction" to the TAUM 1973 report signals a slight shift in emphasis or interest for TAUM, and highlights for readers just how crucial linguistics research was to the work of the team. Text analysis demanded detailed, complex grammars that expressed every occurrence and every possible combination for each word category,[47] and the team was grateful to have obtained

such grammars in French and English from other departments at the University of Montreal and elsewhere.



The cover of TAUM 1971 runs the sentence "You can't see with your eyes closed" through Q-system phases: –01– VOUS + N + ETES + PAS + CAPABLES + DE + VOIR + SI + VOUS + FERM IEZ + VOS + YEUX –02–

Weather Translation: MÉTÉO

Harris recalls Marcel Lacourcière, Assistant to the Translation Bureau Superintendent, suggesting at one point that they focus — like the Americans had — on technical texts, instead of the "anything goes" challenge of translating Hansard. Another branch of government was taxed with producing 37,000 words per day, in the form of short-burst weather bulletins that each had a shelf-life of only four hours. This was a tedious, repetitive job and translator turnover was high.

Esteem tended to rest with loftier and so-called "literary" prose of the debate proceedings, and the most well-trained and longstanding translators held these valued positions. The documentary value of value of weather report translation may have been undervalued in terms of job satisfaction, but this lowballing might be considered inversely proportional to the value of accuracy and regularity in just these kinds of throwaway documents. One intuitively knows, without gathering any data, whether Hansard or weather bulletins are consulted by more Canadians on a daily basis.

The idea of using MT for weather bulletins was first discussed in fall of 1974. Official languages legislation made it mandatory for the Environment Minister to ensure that weather bulletins were available in both official languages everywhere in the country. Bulletins in French were conspicuously lacking. The TB already operated a translation office in Dorval on behalf of the Environment Minister that pertained to forecasts. Rather than bear the costs of training still personnel for a job no one loved doing, the TB considered MT. It was apparently a translator from the Meteorological sub-section who made the initial suggestion.[48]

A prototype that used the Q-systems entered into operation in May 1977 as TAUM-MÉTÉO, "a fully automatic system for the translation of public weather forecasts from English into French covering the whole of Canada."[49] Environmental data collected at various measurement stations was sent to one of eight regional weather centres, then to Toronto via the Canadian National-Canadian Pacific Telecommunications network.[50] Meteorological bulletins were channeled[51] further to a computer (at the time, one of the most powerful in Canada, a Control Data CDC 7600) at the Centre météorologique canadien/Canadian Meteorological Centre in Dorval, a suburb of Montreal. MÉTÉO was part of a communications network that was already in place:

> Whether translated mechanically or manually, from the moment the forecast is typed on the terminal at its weather office of origin, it is available in English and later in French, but only in an electronic form. It does not appear on paper until the very end of the process, at the telexes of the network's subscribers.[52]

English bulletins were first run through a specially designed editor that scanned automatically for untranslatable sentences. These were sentences containing mismatches between human input and system dictionaries, generally words unfamiliar to the system or orthographic errors (e.g. letter "l" in place of number "1"). These were marked for post-editing, and telexed to the Translation Bureau in Ville St- Laurent for human translation. Over 50% of bulletins were rejected in the development stage, but system refinements reduced the castoffs to 20%, a suitable amount, given overall decreases in translation processing times and expenditure. Completed messages, automatic and human translated, were fed back into the CN/CP communications network for distribution to radio, television and print. For over two decades, the system formed (indeed, still forms, in an updated, hybridized variety) one node of a self-monitoring conduit that registered the nation's meteorological pulse every four hours. Chance of accumulating Canada's climatic state of mind: 100%.

MÉTÉO employed a transfer approach to MT, which addressed lexical ambiguities of English input using syntactic and some semantic analysis. This processing readied the data for transfer into French according to program grammars and rewriting rules. The program's linguistic data consisted of three bilingual dictionaries containing "idioms," place names and meteorological vocabulary (a total of 1200 items). Of three processing modules, one performed syntactic analysis of English, and two handled syntactic and morphological generation of French.[53] Due to the domain constraints, MÉTÉO could operate more efficiently without English morphological analysis and the English-French transfer phases, and so represented a simplification of the previous TAUM systems.

The focus on Source Language representation made transfer-based MT a rather unilingual endeavour, in MÉTÉO's case, focusing on English syntax and semantics, with less attention to French. However, this monolingual aspect was displaced somewhat by detecting and articulating an entirely different kind of language in the development process. Richard Kittredge identified the rote, inscriptive formalisms that articulated the authoring of meteorological bulletins as a "sublanguage." The strict formatting, telegraphic writing, lack of verb tenses and articles suggested the team was dealing with an entirely *new* grammar, such that the patterns that characterize the weather bulletin sublanguage do "NOT correspond to the syntactic classes of general English."[54] Rather, they reflect the specific structures, relations and category divisions of a microcosmos that is, ironically, of global breadth, namely, "the world of meteorological observation,"[55] arguably held in common by both English and French.

With respect to time savings, it took an average of 3.8 minutes to generate a machine-translated bulletin as opposed to 30-40 minutes per human-translated bulletin.[56] The total cost in 1976 was one third of the cost of human translation, or 3.5 cents per word.[57] Fifteen years later, the cost had decreased to half a cent per word, with revision affecting only 4% of total output, estimated at 16 million words per year (45,000 words per day), the equivalent of 30 years of human translation.[58]

Unfortunately the conjunction of microworlds that held for weather forecasts did not extend the sublanguage of maintenance manuals for aircraft hydraulic systems. When the Canadian Translation Bureau commissioned the TAUM team in 1976 to develop such a system to handle English-French translation of technical documentation for the Canadian Forces' recently acquired English CP- 140 Aurora maritime patrol aircraft,[59] the task proved significantly more complex. TAUM-AVIATION demonstrated a promising prototype in 1979, but an independent evaluation a year later deemed the system too costly to adapt to the aircraft's full range of subdomains.[60] While the appraisal rating was quite high, the system output simply failed to process 34% of phrases and headings. Interestingly, and arising from the "perfectionism" of MÉTÉO, the system was built to not translate that which it knew it could not — mainly due to words not being part of the dictionary.

Because the system focused only on hydraulics manuals, it would need to develop its dictionaries to handle other areas. However, each English entry took 45 minutes and $10 to create, while each

English-French translation entry took three hours, at $39. In the cost-benefit analysis, it turned out that the number of words wasn't the problem (the manuals seemed formidable, but only contained 100,000 words), but translation requests. Would there be enough documents corresponding to *these* dictionaries awaiting automatic processing to make the venture profitable over the long-term? Moreover, would they correspond in style and vocabulary? Would future manuals match MT system specs?

Canada was having its ALPAC moment. Not wanting to squander its Canadian expertise, it planned to keep a Canadian-grown "third generation MT prototype using artificial intelligence" in mind as a long-term goal. In the meantime, it redirected investment towards developing software tools to boost translator efficiency and started looking at bigger systems coming on the market from the US and Europe. TAUM disbanded in September 1981. TAUM-MÉTÉO, on the other hand, remained, like most translators, in ancillary condition, translating between 9 and 10,000 words per day, three million words per year. Strictly speaking, it was the only MT system in operation in the world.

Medieval Disputation

A decade later, in 1992, Montreal played host to the Fourth International Conference on Theoretical and Methodological Issues in MT. The "statistical turn" in MT was just starting to take hold,[61] so the organizers — including Pierre Isabelle, who'd been a member of the TAUM-MÉTÉO team — selected the apposite theme: "Empiricist versus Rationalist Methods in MT." Brown had summed up the difference bluntly, stating that "rationalist systems are based on information cajoled fact by reluctant fact from the minds of human experts; empiricist systems are based on information gathered wholesale from data.[62] Rationalists were the theoretical linguists (the perfectionists), who adopted rule-based approaches to develop "toy" Artificial Intelligence projects, like MÉTÉO. Empiricists (the engineers) relied on large corpora of translated material. Robert Mercer from the Candide project gave an invited talk — *Rationalist MT: Another Cargo Cult or just Plain Snake Oil?* — and was quoted as saying that "rationalist methods in MT will be on the scrapheap five years from now."[63] As long as there was access to large corpora (by no means assured), triumphs in empiricist MT would follow.

Yorick Wilks, one of the few audible oppositional voices at the conference (from the pragmatist camp more so than the strictly rationalist, he claimed) remarked in turn with regard to IBM's brute force approach: "what they're doing at IBM is not MT: it's an MT factory."[64] He later pointed out that the so-called anti-rationalist position it and its advocates assumed was rhetorical and overstated, reminding colleagues that every MT undertaking involved empiricist and rationalist elements,[65] as well as noting that IBM, in the meantime, had added "rationalist" components to its MT systems in any case. His point was that there was an upper limit to the efficacy of statistics-only methods in MT. The statistical approach could shake philosophical foundations, but could not decisively uproot them.

The organizers put the stand-off between rationalism and empiricism to playful debate during the final session of the conference. In what they termed a "medieval disputation," two advocates from each methodological camp was invited to present arguments, but in favour of the opposite side. The audience was encouraged to heckle and jeer as desired, or as necessary.

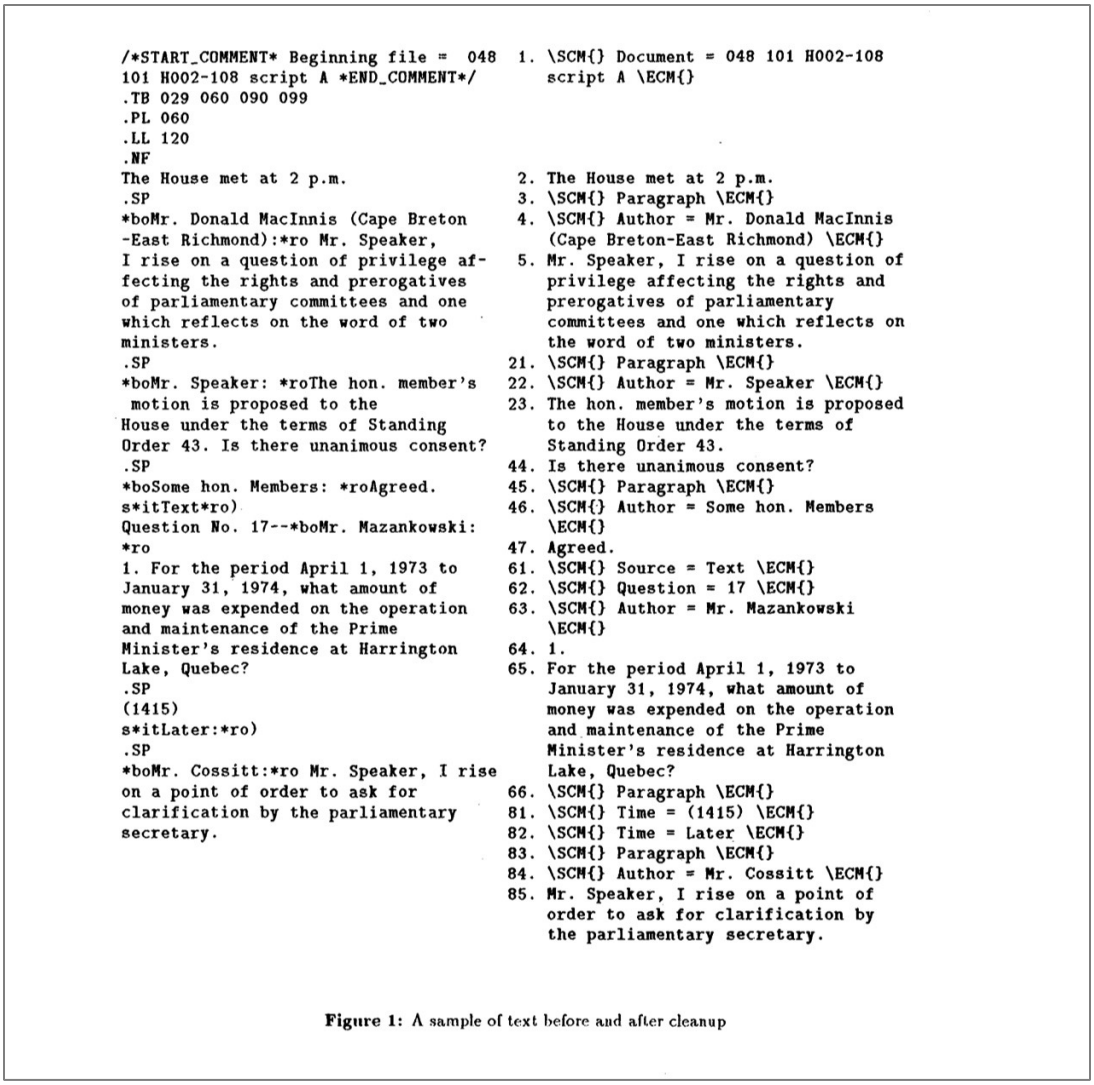
In defense of empiricism were claims about mathematics being grounded in the physical world — the same place the translator's skill resides. Statistical techniques didn't trade in abstract symbols and transformational rules that translators wouldn't even recognize. While empirical methods could never be completely right, at least they weren't based on flaky intuitions. Empirical methods were at last able to address the difference in scale between the complexity of language and the limitations of the human brain. Brains could "do" language, but could only know so much about it (at any one time).

A point made in favour of rationalism was the obvious fact that humans had knowledge and experience that they applied all the time. Extreme empiricism claimed it operated from some place beyond it. Didn't empiricism presume a kind of cognitive rationalism? Ken Church capped off the debate by characterizing the statistical approach as a "free lunch," pinning a critique of his own empiricist position to value and accuracy: "easy answers don't cost too much… their wrong answers are pretty cheap."

One point both camps held in common was that human language did not need to be learned. What none of the researchers addressed was whether *translation* needed to be.

Translation is Pairs of Sentences

Peter Brown and Bob Mercer of the IBM Candide project recalled that colleague and mentor John Cocke learned of the existence of Hansard by chance on an airplane — and most likely during an inebriated conversation. This was no trivial discovery in 1987, as the corpus was massive and, most importantly, machine-readable. It was acquired on magnetic tape and in an "obscure text markup language." Brown and Mercer attributed their switch from speech recognition to MT *directly* to the availability of Hansard data. It was Cocke who had suggested that something might be learned about translation by looking at French and English side by side. For their Fundamental Equation to work, they needed to build a language model for English, based on any English data. But to build a probability model of the translation process, they needed a very special corpus, specifically, "we need pairs of sentences that are translations of one another."[66] Crucially, this parallelism both represented *and* circumvented meaning. As Mercer explained: "The Hansards data […] is a great place in which to investigate meaning […] The French that's written there is the meaning of the English that's written in the other place. You don't need to worry about all that intermediate stuff of what it really means."[67]



```
/*START_COMMENT* Beginning file =  048
101 H002-108 script A *END_COMMENT*/
.TB 029 060 090 099
.PL 060
.LL 120
.NF
The House met at 2 p.m.
.SP
*boMr. Donald MacInnis (Cape Breton
-East Richmond):*ro Mr. Speaker,
I rise on a question of privilege af-
fecting the rights and prerogatives
of parliamentary committees and one
which reflects on the word of two
ministers.
.SP
*boMr. Speaker: *roThe hon. member's
 motion is proposed to the
House under the terms of Standing
Order 43. Is there unanimous consent?
.SP
*boSome hon. Members: *roAgreed.
s*itText*ro)
Question No. 17--*boMr. Mazankowski:
*ro
1. For the period April 1, 1973 to
January 31, 1974, what amount of
money was expended on the operation
and maintenance of the Prime
Minister's residence at Harrington
Lake, Quebec?
.SP
(1415)
s*itLater:*ro)
.SP
*boMr. Cossitt:*ro Mr. Speaker, I rise
on a point of order to ask for
clarification by the parliamentary
secretary.
```

```
1. \SCM{} Document = 048 101 H002-108
   script A \ECM{}
2. The House met at 2 p.m.
3. \SCM{} Paragraph \ECM{}
4. \SCM{} Author = Mr. Donald MacInnis
   (Cape Breton-East Richmond) \ECM{}
5. Mr. Speaker, I rise on a question of
   privilege affecting the rights and
   prerogatives of parliamentary
   committees and one which reflects on
   the word of two ministers.
21. \SCM{} Paragraph \ECM{}
22. \SCM{} Author = Mr. Speaker \ECM{}
23. The hon. member's motion is proposed
   to the House under the terms of
   Standing Order 43.
44. Is there unanimous consent?
45. \SCM{} Paragraph \ECM{}
46. \SCM{} Author = Some hon. Members
   \ECM{}
47. Agreed.
61. \SCM{} Source = Text \ECM{}
62. \SCM{} Question = 17 \ECM{}
63. \SCM{} Author = Mr. Mazankowski
   \ECM{}
64. 1.
65. For the period April 1, 1973 to
   January 31, 1974, what amount of
   money was expended on the operation
   and maintenance of the Prime
   Minister's residence at Harrington
   Lake, Quebec?
66. \SCM{} Paragraph \ECM{}
81. \SCM{} Time = (1415) \ECM{}
82. \SCM{} Time = Later \ECM{}
83. \SCM{} Paragraph \ECM{}
84. \SCM{} Author = Mr. Cossitt \ECM{}
85. Mr. Speaker, I rise on a point of
   order to ask for clarification by
   the parliamentary secretary.
```

Figure 1: A sample of text before and after cleanup

The left column is the original Hansard data, in an "obscure markup language." (Brown, 1988). Right column is IBM's cleaned up text, which retains most Hansard headings and subheadings.

Unsuccessful MT development had been obsessed by this "intermediate stuff," overwhelmed by intuition, and stuck too closely to the idea that "a translator proceeds by first understanding the French, and then expressing in English the meaning that he has thus grasped."[68] IBM's feat was to rearticulate the translation process as one void of understanding, supplanting an interpretive and expressive act of re-inscription with statistical probabilities. From this new point of view, the formal point of view, the equation was "completely adequate." (More, however, could be said about the methods IBM devised to classify as successful results that were more *and* less correct.)

What makes Hansard the perfect corpus, therefore, is this simple concept of alignment, of one language mirroring another. That which Lydia Liu might call its "hypothetical equivalence"[69] is a product of the Canadian government's institutional commitment to language parity, to voicing

English and French simultaneously, as expressed not only through Hansard, but through the mechanism of its Translation Bureau and across all of its products, including something as forgettable and persistent as MÉTÉO.

Candide definitively altered the course of MT research and its approach forms the backbone of most of today's MT software. Brown et al make some interesting conjectures in early papers, just before they left the field of computational linguistics to found Renaissance Technologies, a hedge-fund firm that uses statistical modelling to seek whisper-quiet correlations in data to guide financial investments:

> Our work has been confined to French and English, but we believe that this is purely adventitious: had the early Canadian trappers been Manchurians later to be outnumbered by swarms of *conquistadores,* and had the two cultures clung stubbornly each to its native tongue, we should now be aligning Spanish and Chinese.
>
> We conjecture that local alignment of the component parts of any corpus of parallel texts is inherent in the corpus itself, provided only that it be large enough. Between any pair of languages where mutual translation is important enough that the rate of accumulation of translated examples sufficiently exceeds the rate of mutation of the languages involved, there must eventually arise such a corpus.[70]

Facts about parallel text correlation that we can basically agree upon start to look less convincing when they're used to make claims about corpus production and maintenance as a neutral function or obvious outcome of culture clash or community relation. Aligned parallel corpuses like the Hansard, and of that size, are quite rare objects. One wonders what might have become of Kathleen Booth's statistical analyses, had her research contract been renewed and had she had access to more powerful machines. The surest thing to emerge so far from empiricist MT is post-editing. It's important to consider the material conditions, the values and priorities, the routines, training regimens and technologies that facilitated their production and manifested *these* particular alignments, and left out others, as well as the work these alignments are (and were) expected to do in their specific contexts of use.

Around the same time C.B. Watt won his lobby for MT research, it was suggested that computerized printing be adopted for Hansard and other parliamentary publications. Replacing the hot-metal system then in use with new technology would help reduce quickly mounting costs in connection with the need to print everything in French and English. The system included: a UNIVAC 1050 computer, a teletypesetting system, paper tape-producing keyboards and paper tape-driven Elektron line-casters, as well as special text-editing program modified to the particular operating needs of the House of Commons. Alexander Small, Second Clerk Assistant of the House of Commons, described the multi-year system upgrade in a 1974 article, which also foretells of provisions for "future potential uses" of machine-readable bilingual databases as language-translation aids once the system allowed for automatic information retrieval. One detail

now seems plausible only if severed from the context above, the fact that these were "within reach almost as soon as published […] by-products at little or no extra cost":

> The first broad objective is to publish all parliamentary publications bilingually by computerized processes which produce a simultaneous data-base as a by-product at little or no extra cost in both official languages thereby bringing a second broad objective of automated information retrieval within reach almost as soon as published and, as a third broad objective providing, as another by-product, machine-readable data bases of all English and French texts printed in any parliamentary publication […] for computerized searches to aid in translation to and/or from the two official languages.[71]

[1] .992 probability for *bravo* as opposed to .005 for *entendre.* Peter J. Brown et al, "A Statistical Approach to Language Translation," *Computational Linguistics*, 16(2) (1990), 83.

[2] Brown et al., "A Statistical Approach," 83.

[3] Adam L. Berger et al, "The Candide System for Machine Translation." *Human Language Technology: proceedings of a workshop held at Plainsboro, New Jersey, USA, March 8-11, 1994* (San Mateo, CA: Morgan Kaufmann, 1994): 157-162.

[4] Berger et al, "The Candide System," 157.

[5] Brown et al., "A Statistical Approach," 82.

[6] Brown et al, "A Statistical Approach," 79.

[7] Brown, Peter E, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. "The mathematics of statistical machine translation: Parameter estimation." Computational Linguistics 19(2) (1993): 265.

[8] Expressed as a "fundamental equation," the method looks for a sentence in language E (IBM used English) that is most probable given a particular sentence in language F (IBM used French). Bayesian reasoning is used to calculate probabilities associated with the source (statistical details particular to language E, also called "source modelling") and probabilities associated with the channel (statistical details particular to translations between E and F, also called "channel modelling"). Multiplying the two expresses the probability of (e) given (f). Kevin Knight explains it this way in a tutorial workbook on statistical MT: "the most likely translation (e) maximizes the product of two terms (1) the chance that someone would say (e) in the first place and (2) if he did say (e) the chance that someone would translate it into (f)." Kevin Knight, "A Statistical MT Tutorial Workbook," (1999): http://www.isi.edu/~knight/.

[9] Yorick Wilks, "Stone Soup and the French Room," Current issues in computational linguistics: in honour of Don Walker, eds. Zampolli, A., N. Calzolari & M. Palmer (Linguistica Computazionale, vol. 9-10), Pisa, Dordrecht (1994): 591.

[10] David E. Wellbery, foreword to *Discourse Networks 1800/1900* by Friedrich Kittler (Stanford: Stanford University Press, 1990), xii. The work of Friedrich Kittler was especially influential and focused on the materialities of communication from a post-hermeneutic perspective. Attention is given to inscriptive channels, but in the form of "technologies and institutions that allow a given culture to select, store, and process relevant data." Friedrich Kittler, *Discourse Networks 1800/1900* (Stanford: Stanford University Press, 1990), 369. As David Wellbery describes it: "The thrust of Kittler's analysis is to show that as long as we continue to operate within the hermeneutic paradigm we are paying homage to a form of language processing long since deceased."

[11] Lawrence Venuti, The Translator's Invisibility: A History of Translation (New York & London: Routledge), 1995; Jay David Bolter and Richard Grusin, Remediation. (Cambridge: MIT Press), 1998.

[12] See, for instance, Wendy Chun, "On Software, or the Persistence of Visual Knowledge" Grey Room Winter 2005, No. 18: 26–51; N. Katherine Hayles, My Mother Was a Computer: Digital Subjects and Literary Texts (Chicago: Chicago University Press): 2005; Leah Price and Pamela Thurschwell (eds.), Literary Secretaries/Secretarial Culture (Aldershot and Burlington: Ashgate): 2005.

[13] Clough, P., K. Gregory, B. Haber, and J. Scannell. 2014. "The Datalogical Turn." In *Nonrepresentational Methodologies: Re-Envisioning Research*, ed. Phillip Vannini. Oxford: Taylor & Francis; dana boyd and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, communication & society* 15.5 (2012): 662–679.

[14] John Hutchins, "ALPAC: the (in)famous report," *MT News International* 14 (June 1996), 15-21.

[15] John R. Pierce, B. Carroll, et al., *Language and Machines — Computers in Translation and Linguistics*. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC, 1966.

[16] See, for instance, Michael D. Gordin, *Scientific Babel: How Science Was Done Before and After Global English* (Chicago: University of Chicago Press): 2015; Jacqueline Léon, Histoire de l'automatisation des sciences du langage (Lyon: ENS Éditions): 2015; David Nofre, Mark Priestley and Gerard Alberts, "When Technology Became Language: The Origins of the Linguistic Conception of Computer Programming, 1950–1960," Technology and Culture, Volume 55, Number 1, January 2014, pp. 40-75.

[17] John R. Pierce, B. Carroll, et al., *Language and Machines — Computers in Translation and Linguistics*. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC, 1966.

[18] Elliott Macklovitch "MT R&D in Canada," MT Summit VI. Machine Translation: Past, Present, Future. Proceedings (San Diego, California, 1997): 200.

[19] W. John Hutchins, "A new era in machine translation research," Translating and the Computer 16. [Paper presented at a conference... 10-11 November 1994, Institution of Civil Engineers, London SW1] (London: Aslib, 1994): 11.

[20] Letter of 27 July, 1965. Library and Archives Canada. RG-77, 1981-82/031. Box 12, File 5230-4-5.

[21] Brown doesn't recollect the demo, but doubts he would have described it in these terms. Personal communication.

[22] Letter of June 19, 1963. Library and Archives Canada. RG-55, Vol. 278. TB 612588 (June 20, 1963).

[23] Letter of June 19, 1963. Library and Archives Canada. RG-55, Vol. 278. TB 612588 (June 20, 1963).

[24] Equal status is accorded to French and English in the House of Commons, though French was rarely spoken in the House until the late 1950s.

[25] Alexander Small, "Bilingual Publishing and Retrieval of Parliamentary Publications in the House of Commons of Canada from On-Site Terminals to Remote Off-Site Computerized Photocomposition-Typesetting Processes." (Article for 1974 volume of *The Table*.) April 17, 1974.

[26] "Miscellaneous." 612588, National Research Council. Feasibility Study into Possibility of Translation from English to French by Computer Methods. Library and Archives Canada. RG-55, Vol. 278. TB 612588 (June 20, 1963).

[27] See W. John Hutchins, "The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954," 1. Unpublished, available at http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf. IBM Press release (1954). "701 Translator," January 8, 1954. http://www- 03.ibm.com/ibm/history/exhibits/701/701_translator.html.

[28] House of Commons Debates, 26th Parliament, Session 2, Volume 12. March 19, 1965, 12581.

[29] Letter of July 22, 1965.

[30] Margaret Masterman, "Man-aided computer translation from English into French using an on-line system to manipulate a bi-lingual conceptual dictionary, or thesaurus," *Coling 1967: Conference Internationale sur la Traitement Automatique des Langues* (Grenoble, France, August 1967), 4.

[31] Andrew D. Booth and Kathleen H.V. Booth, "The Beginnings of MT." In W. John Hutchins (ed.) Early Years in Machine Translation: Memoirs and Biographies of Pioneers. (Amsterdam/ Philadelphia: John Benjamins), 257.

[32] Booth and Booth, "The Beginnings of MT." 257.

[33] Kathleen H.V. Booth, "Machine aided translation with a post-editor." In Andrew D. Booth (ed.) *Machine Translation* (Amsterdam: North-Holland Publishing Company, 1967), 51-76.

[34] Andrew D. Booth, "Introduction," in *Machine Translation,* ed. Andrew D. Booth. (Amsterdam: North-Holland Publishing Company, 1967), ix.

[35] Andrew D. Booth, "Mechanical Resolution of Linguistic Problems," in *Electronic Information Handling*, ed. A. Kent & O.E.Taulbee. (Washington, DC: Spartan Books, 1965), 47.

[36] K.H.V. Booth, C.A. Brown, C. Stock, "English-French Translation on a Computer," *Pensiero e linguaggio in operazioni, Thought and language in operations*, Vol. 2, 1971, 231.

[37] Booth, Brown & Stock, "English-French Translation," 233.

[38] Booth, Brown & Stock, "English-French Translation," 233-234.

[39] Personal communication.

[40] CETADOL Quarterly Report. October 1967.

[41] Dugas writes, "*parce qu'actuellement un situation politique prévaut en ce sens au Canada et convient à la majorité des lecteurs francophones à qui s'adresse cette traduction."*

[42] Colmerauer had been discharging his French military service at the University of Montreal since 1967.

[43] Q-systems is considered the precursor to Prolog, a logic programming language for Artificial Intelligence and computational linguistics developed by Colmerauer upon his return to France in 1972.

[44] TAUM 1971, Université de Montréal.

[45] Alain Colmerauer and Philippe Roussel, The Birth of Prolog, 1999. Draft of paper in History of Programming Languages, edited by Thomas J. Bergin and Richard G. Gibson, ACM Press/ Addison-Wesley, 1996. http://alain.colmerauer.free.fr/alcol/ArchivesPublications/PrologHistory/19november92.pdf

[46] TAUM 1971, 4.

[47] Kittredge: *présuppose une grammaire qui exprime toutes les possibilités et combinaisons d'occurrences pour chaque categorie du mot. TAUM 1973, Montreal.*

[48] Benoît Thouin, "The METEO System." In V. Lawson (ed.), Practical Experience of Machine Translation. (North-Holland Publishing Company: ASLIB, 1982), 43.

[49] John Chandioux, "METEO, an operational system for the translation of public weather forecasts," [FBIS Seminar on Machine Translation, 8-9 March 1976, Rosslyn, Virginia]. American Journal of Computational Linguistics, microfiche 46: 27.

[50] John Chandioux, "MÉTÉO: un système opérationnel," 127.

[51] John Chandioux, "MÉTÉO: un système opérationnel," 127. Chandioux writes "*extraites du flot et canalisées*."

[52] Thouin, "The METEO System," 39.

[53] W.John Hutchins and Harold L. Somers. *An introduction to machine translation*. (London: Academic Press, 1992), 209-210.

[54] Richard I. Kittredge, "The Significance of Sublanguage for Automatic Translation," *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages* (Colgate University, Hamilton, New York, August 14-16, 1985), 157.

[55] Kittredge, "The Significance of Sublanguage," 157.

[56] Chandioux & Guéraud, "MÉTÉO: un système à l'épreuve du temps," *Meta*, 26(1) (1981): 21.

[57] Jean Chandioux, "METEO, an operational system for the translation of public weather forecasts," FBIS Seminar on Machine Translation, March 8-9, 1976, Rosslyn, Virginia. American Journal of Computational Linguistics, microfiche 46, 28, 36.

[58] Jean Chandioux, "Meteo: Environment Canada," *The MT user experience. MT Summit III. Proceedings* (July 1-4, 1991, Washington, DC): 123.

[59] Hutchins, *Past, Present, Future*, 231; John Chandioux, "Creation of a Second-Generation System for Machine Translation of Technical Manuals," *Overcoming the language barrier*, 3- 6 May 1977, vol. 1. (Munchen: Verlag Dokumentation): 614.

[60] Pierre Isabelle & Laurent Bourbeau, "TAUM-AVIATION: its technical features and some experimental results," *Computational Linguistics*, Volume 11(1) (1985): 18.

[61] Yorick Wilks, "Some notes on the state of the art: Where are we now in MT: what works and what doesn't? And the role MT as an international collaborative activity," *International conference "Machine translation: ten years on" Proceedings ... held at Cranfield University, England, 12-14 November 1994* (Cranfield University Press, 1998): 1.

[62] Peter F. Brown et al, "But dictionaries are data too." *Human Language Technology: proceedings of a workshop held at Plainsboro, New Jersey, USA, March 21-24, 1993* (San Francisco, CA: Morgan Kaufmann, 1993): 202.

[63] Yorick Wilks, "Stone Soup and the French Room," Current issues in computational linguistics: in honour of Don Walker, eds. Zampolli, A., N. Calzolari & M. Palmer (Linguistica Computazionale, vol. 9-10), Pisa, Dordrecht (1994): 591.

[64] Wilks, "Stone Soup," 591. Wilks was working on a competing system (Pangloss), but it is worth noting that both Pangloss and IBM Candide (and one other system) were all carried out under the auspieces of a single DARPA-funded project (Frederick Jelinek, "Some of My Best Friends are Linguists," *Language Resources and Evaluation*, 39 (2005): 32).

[65] Insofar as the MT is defined by the achievement of some sort of 'success,' Wilks writes, "we are all empiricists and, to a similar degree, we are all rationalists, in that we prefer certain methodologies to others and will lapse back to others only when our empiricism forces us to" Wilks, "Stone Soup," 591.

[66] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. "A Statistical Approach to Machine Translation," Computational Linguistics, 16(2), June 1990, 82.

[67] Peter Brown & Bob Mercer, "Oh, yes, everything's right on schedule, Fred." Twenty Years of Bitext. http://cs.jhu.edu/~post/bitext/

[68] Brown et al, "The Mathematics," 265.

[69] Lydia Liu, ed., Tokens of Exchange: The Problem of Translation in Global Circulations (Durham, Duke University Press, 1999). “In order for the process of circulation to take place at all, the agents of translation on each side start out by hypothesizing an exchange of equivalent meanings, even if the hypothesis itself is born of a structure of unequal exchange and linguistic currency.” (21) Further: “The act of translation thus hypothesizes an exchange of equivalent signs and makes up that equivalence where there is none perceived as such.” (34)

[70] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, “The Mathematics of Statistical Machine Translation: Parameter Estimation,” Computational Linguistics 18(2), 296.

[71] Alex Small, Article on Computerization of House of Commons Publications for 1974 Volume of “The Table.” Ottawa, April 17, 1974. (File CTCP-TS-1)