Obituary

Victor H. Yngve

W. John Hutchins*

Victor Yngve (5 July 1920 to 15 January 2012) was a major contributor in a number of fields within computational linguistics: as the leading researcher in machine translation (MT) at the Massachusetts Institute of Technology (MIT), as editor of its first journal, as designer and developer of the first non-numerical programming language (COMIT), and as an influential contributor to linguistic theory.

While still completing his Ph.D. on cosmic ray physics at the University of Chicago during 1950–1953, Yngve had an idea for using the newly invented computers to translate languages. He contemplated building a translation machine based on simple dictionary lookup. At this time he knew nothing of the earlier speculations of Warren Weaver and others (Hutchins 1997). Then during a visit to Claude Shannon at Bell Telephone Laboratories in early 1952 he heard about a conference on machine translation to be held at MIT in June of that year. He attended the opening public meeting and participated in conference discussions, and then, after Bar-Hillel's departure from MIT, he was appointed in July 1953 by Jerome Wiesner at the Research Laboratory for Electronics (RLE) to lead the MT research effort there. (For a retrospective survey of his MT research activities see Yngve [2000].)

Yngve, along with many others at the time, deprecated the premature publicity around the Georgetown–IBM system demonstrated in January 1954. Yngve was appalled to see research of such a limited nature reported in newspapers; his background in physics required experiments to be carefully planned, with their assumptions made plain, and properly tested and reviewed by other researchers. He was determined to set the new field of MT on a proper scientific course. The first step was a journal for the field, to be named *Mechanical Translation*—the field became "machine translation" in later years. He found a collaborator for the journal in William N. Locke of the MIT Modern Languages department. The aim was to provide a forum for information about what research was going on in the form of abstracts, and then for peer-reviewed articles. The first issue appeared in March 1954.

Yngve's first experiments at MIT in October 1953 were an implementation of his earlier ideas on word-for-word translation. The results of translating from German were published in the collection edited by Locke and Booth (Yngve 1955b). One example of output began:

> Die CONVINCINGe CRITIQUE des CLASSICALen IDEA-OF-PROBABILITY IS eine der REMARKABLEen WORKS des AUTHORs. Er HAS BOTHen LAWe der GREATen NUMBERen ein DOUBLEes TO SHOWen: (1) wie sie IN seinem SYSTEM TO INTERPRETen ARE, (2) THAT sie THROUGH THISe INTERPRETATION NOT den CHARACTER von NOT-TRIVIALen DEMONSTRABLE PROPOSITIONen LOSen ...

* Previously at the University of East Anglia, Norwich, UK. E-mail: john@hutchinsweb.me.uk.

It was obvious that the output was poor, but nevertheless it appeared to be good enough for scientists with some knowledge of German grammar to read and extract relevant information. Yngve concluded that word-for-word translation could be taken as a first approximation. But a major problem was that many words have more than one meaning out of context.

Erwin Reifler had suggested the use of pre-editors who would annotate texts before translation; and Victor Oswald proposed the use of microglossaries (dictionaries limited to one specific narrow field) in order to reduce the number of homographs. But Yngve believed that the problem of multiple meanings could be resolved with syntactic analysis. Nonetheless, he rejected Bar-Hillel's "operational syntax" put forward at the 1952 conference (and later known as categorical syntax) as too complex for long sentences and too remote from traditional grammatical distinctions. He had been impressed by the work of Bloomfield and Fries, and had become convinced that linguistics could provide stable and repeatable methods akin to the procedures of cosmic ray physics. Therefore, over the following years, he appointed a number of linguists to the RLE staff, starting with Noam Chomsky in 1955, in order to undertake basic research. He was disappointed, however, that most of the linguists he hired were more interested in pursuing their theoretical studies than in tackling the practical problems of MT.

Yngve's approach to syntactic analysis was to begin with the identification and isolation of the different possible grammatical functions of words. The aim was to set up mutually exclusive word-classes (i.e., one class for the noun function, another for the verb function, and so on) The approach was to set up substitution frames to isolate contexts. Thus *walk* may occur in the frame

(1) The ____ was enjoyable

or in the frame

(2) They ____ home every day.

Taking words from a corpus and testing in each frame produces a matrix of different contexts (substitution frames) and the words occurring in those contexts. These words form a word-class and their frames form context classes. As a result, each sentence sequence of word-classes would determine a unique sequence of context classes. An algorithm was proposed which searched left-to-right for the longest match for a sequence of word-classes. Sequences of word-classes formed phrases and clauses, so the algorithm was capable of looking also for phrase sequences. On this basis, a table-driven algorithm for syntactic analysis was designed and implemented (Yngve 1955b). This demonstrated for the first time in MT the importance and practical value of separating descriptive linguistic information (in this case, words and contexts) and language-independent algorithmic procedures (searching and matching). The practice was widely adopted by other groups in MT and in computational linguistics.

At the same time, work was going on at the RLE on investigations of coding and switching theory and on Shannon's information theory. Yngve decided to investigate error correcting properties of language (Yngve 1954). Sequences of letters and words are not random but constrained by codes specifying which sequences are legitimate. Testing a text for deviations from randomness would help to reveal its structure. A first step would be the location of all occurrences of a given frequent word and the determination of its effect on the occurrence of other frequent words in the neighborhood by comparing their frequency of occurrence with what would be expected if they occurred at random. For this investigation of what was to be called "gap analysis" (Yngve 1956)

a computational experiment was conducted on a corpus of just 9,490 words. First, the most frequent words were identified: *the* (599 instances), *to* (252), *of* (241), *a* (221), *and* (207), and *in* (162). Then, the gaps between these words were determined (in numbers of intervening words). In the case of *of* and *the* the gaps were one (*the* ___ *of*), two (*the* ___ ___ *of*), three (*the* ___ ___ ___ *of*), and so on, with frequencies of 72, 31, and 6, respectively. These results pointed to syntactic constraints on constructions with *of* and *the*. Further results indicated that "structures with *the* can be expected to have two or three words, and constructions with *and* frequently involve at least fifteen words." Similar observations were that "*of* is different from *to* in that it frequently follows *a* or *the* with a gap of one or two." And so forth, in a procedure which has now become familiar in statistics-based computational linguistics. Yngve was a pioneer. Unfortunately, at the time, these encouraging results could not be pursued because of the lack of large enough corpora in machine-readable form.

The parallels between coding theory and linguistics suggested a two-stage model for machine translation, where speakers encode a message that is decoded by a recipient (Yngve 1955a). Between encoding and decoding there would be some representation of the 'message' to be preserved in translation. Further consideration recognized that structural representations of input would not be the same as those required for output, however. Attention, therefore, focused on the transition stage where the meaning or message of the input language would be expressed and where choices would be made for producing output sentence structure. Thus the MIT group saw the need for a three-stage model: syntactic analysis, structure transfer, and synthesis—a model for many transfer-based MT systems in subsequent years. Yngve's model was, however, purely syntactic; no attempt was made at this time to include semantic information (Yngve 1957).

At the same time as the development of the transfer model, Yngve and his colleagues tackled the problem of how linguists could be productive in MT research. They decided that what was needed was a programming system that accepted notations of a kind familiar to linguists, rather than systems such as FORTRAN and COBOL designed for mathematicians and business applications. Thus, in collaboration with the MIT Computation Center, they developed the first programming language designed specifically for string-handling and pattern-matching, COMIT (Yngve 1961a, 1967). It was ready for testing in late 1957—thus antedating by two years LISP, another programming language devised for linguistics applications. COMIT was later used as a basis for the SNOBOL language.

The availability of COMIT meant that the MIT group could proceed further with the three-stage model in the development of an algorithm for sentence production. Initially the generative grammar of Chomsky seemed to be an ideal model. What was required in MT, however, was not the generation of all grammatical sentences from a given source but particular specifiable sentences in context. The rules of a grammar derived from a simple children's story were written in COMIT in 1959 and tested in a program of random sentence production (Yngve 1961b). Its main objective, in which it succeeded, was to test the validity of the grammar rules, particularly rules for discontinuous constructions and for coordination.

One outcome of programming the sentence-production algorithm with COMIT was the "depth hypothesis," for which Yngve is probably now best known (Yngve 1960a) both in linguistics and in computational linguistics. The transformational approach had already been rejected because it required too much storage space. The next question was how much storage (push-down store) was needed for discontinuous (left-branching) expansion and for regular (right-branching) expansion. It was clear that right expansion

(or "progressive" application) was potentially infinite: *the dog that worried the cat that killed the rat that ate the malt* .... On the other hand, left expansion (or 'regressive' application) was limited: *the malt that the rat that the cat that the dog worried killed ate*. Yngve calculated the maximum amount of extra temporary memory required to produce a sentence (i.e., the number of unexpanded constituents at any one time [its depth]). He found that in practice even very long sentences seldom had a depth of more than two or three. Sentences with depths of more than three, such as the "regressive" constructions, were often considered ungrammatical and/or unintelligible. Yngve noted the relationship of this linguistic feature to the restrictions on immediate memory and processing identified by Miller (1956). Most languages include mechanisms for restricting the depth of regressive constructions, such as agglutination and compounding.

The depth hypothesis accounted for and predicted many syntactic features of English, including its historical changes, and also appeared to account for many features in other languages. Yngve recognized that it arose from MT research and not from linguistic theory, however. It was a hypothesis that needed to be tested against empirical evidence. Its significance for linguistics was widely recognized from the beginning, but it did not conform to the preconceptions of Chomskyan theory. Although Chomsky had championed the rigorous statement of theory and its strict application to linguistic material without ad hoc adjustments (Chomsky 1957, page 5), he regarded the depth hypothesis not as a testable scientific hypothesis but as a rival linguistic theory. For Yngve this attitude was unscientific.

Throughout his time at MIT Yngve stressed the need for a "basic, long-range approach to translation" and not to look for "short-cut methods that might yield partially adequate translations at an early date" (Yngve 1960b, page 183). No working MT system emerged from MIT, therefore, but the quality of the research is incontestable. Apart from Yngve's own contributions to many aspects of syntactic analysis (the three-stage transfer model, the depth hypothesis, and not least computer programming), his colleagues also made significant contributions in a variety of areas: grammar theory (Gilbert Harman and Kenneth Knowlton), semantics (Elinor Charney), logic and language (Jared Darlington), transfer and interlingua (Edward Klima), computer storage (William S. Cooper), Arabic translation (Arnold Satterthwait), and French translation (David Dinneen). Citations for these contributions will be found in Yngve's comprehensive survey of MT research at MIT (Yngve 1967; see also Yngve 2000).

By 1964, Yngve had come to the conclusion that "work in mechanical translation has come up against what we will call the semantic barrier ... we will only have adequate mechanical translations when the machine can 'understand' what it is translating and this will be a very difficult task indeed" (Yngve 1964, page 279). Understanding involved the background knowledge that people bring to the comprehension of communication. He could see no solutions coming from linguistics. For many years, Yngve had grown increasingly doubtful about the health of linguistic science and consequently about the feasibility of good quality MT in general.

By 1965, funding for the MIT research group had ceased—perhaps in anticipation of the ALPAC report which had a major impact on the funding of all US research groups—and in that year Yngve went back to the University of Chicago as head of the Department of Library Sciences, Linguistics, and Behavioral Sciences.

By this time, the journal *Mechanical Translation*, which he had founded in 1954, was coming to an end. The aim had been to provide a public record of achievement in MT. His ambitions for the journal could not be fulfilled, however, for various reasons. There were relatively few outstanding contributions submitted for publication; many MT researchers were funded by government bodies, which required the submission of

fairly frequent reports, and these reports were distributed widely to other researchers and public institutions. Researchers believed they had fulfilled their duties to publicize research, and did not see the need to submit shorter articles to a peer-reviewed journal which would probably not appear for several months. In addition the journal could not survive solely on subscription charges, and authors were asked to contribute a page charge towards publication costs, which they were reluctant to do.

In June 1962, the Association for Machine Translation and Computational Linguistics was founded, with Yngve as its first president. The inclusion of 'computational linguistics' in the title was indicative of the ever-expanding range of activities in the field of natural language processing; MT was only part, and a diminishing proportion. The Association took over Yngve's journal with a changed title, *Mechanical Translation and Computational Linguistics*, and Yngve remaining as editor. Even with the inclusion of many more articles in computational linguistics and significantly fewer in MT, however, the journal became ever more irregular and it was wound up in 1970. In 1968 machine translation had already been dropped from the Association's title.[1]

From this time on, Yngve turned away from his MT interests to questions of linguistic theory that had been his increasing concern since the publication of his "depth hypothesis." From 1965, Yngve published a series of papers devoted to the foundations of linguistic theory, many at the conferences of LACUS (Linguistic Association of Canada and the United States). Just as when he had founded the *Mechanical Translation* journal in 1954, he was determined to put studies of language on a sound scientific footing. His recurrent theme was the unscientific nature of current linguistics. In this period, he set forth the framework of what he called "human linguistics" (later "hard-science linguistics") where the units being analyzed are not the traditional properties of sentences, verbs, noun phrases, gender, tense, phonemes, and so on, which, as he demonstrated, are unfounded assumptions derived often from Greek philosophical speculations. Instead the basic participants are communicating individuals and physical observable "props" (abstractions of relevant physical objects, clocks, doors), "channels" of communication (sound waves, writing, signs), and "settings" (relevant aspects of the physical environment, ticket counters, rooms). Yngve's first attempt at summarizing and formulating his theory was published in his book, *Linguistics as Science* (Yngve 1986).

Expansion and elaboration of his theoretical standpoint followed in further papers and were brought together in *From Grammar to Science* in 1996. The opening chapters are a cumulative rejection of all traditional and contemporary linguistics and philosophy of language (from the Greeks to Saussure, Bloomfield, Fries, Harris, Chomsky, and many others), including a rejection of his own widely accepted "depth hypothesis." The basic contention is that fundamental concepts of linguistics are intuitions and not based on the observable behavior of people communicating. Yngve describes a detailed comprehensive program for a new foundation of linguistics in which "we abandon the unsupported assumptions of signs, words, meanings, and sentences" and move completely into "the world of sound waves and the people who speak and understand where we can test our theories and hypotheses against observations of the physical reality all the way down to the basic assumptions of all science" (Yngve 1996, page 308). He had not lost sight of computational treatments, and included (chapters 19 and 20) an implementable notation for representing and testing hypotheses.

1 For Yngve's reflections on these changes in MT and computational linguistics see Yngve (1982).

Yngve attracted the support and collaboration of a growing number of scholars sympathetic to his goal of a "hard-science linguistics," and in 2004 a volume of essays appeared dealing with a wide range of issues and topics under this title (Yngve and Wąsik 2004). Yngve's own contributions included an examination of communications in formal meetings and an essay on the foundations of a hard-science phonetics and phonology. Others wrote about speech acts, anaphora, business negotiations, language change, educational discourse, communication in science, and much more. The range of applications is impressive, but it is still too early to say what the future impact of Yngve's "hard science" approach may have in the study of communication in all its forms. He was himself convinced that in this "necessary reconstruction" of linguistics on a truly scientific basis, "computational linguistics is destined to play an essential role" (Yngve 1982, page 94).

Throughout his long career, Vic Yngve retained a modesty about his considerable achievements in machine translation and computational linguistics and a firm commitment to the highest standards of research practice which impressed everyone who knew him. He will probably be best remembered for his depth hypothesis, but his other papers on MT should continue to interest all who look for insights in the computational analysis of natural language—and not just for historical reasons. His articles and books on "hard-science linguistics" deserve to be essential reading for anyone reflecting upon the foundations of linguistics, and indeed anyone concerned with the general health of current and future studies of language and communication.

References

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague.

Hutchins, John. 1997. From first conception to first demonstration: The nascent years of machine translation, 1947–1954. *Machine Translation*, 12:195–252.

Miller, George A. 1956. The magical number seven, plus or minus two: Some limits in our capacity for processing information. *Psychological Review*, 63:81–97.

Yngve, Victor H. 1954. Language as an error correcting code. *Quarterly Progress Report of the Research Laboratory of Electronics*, MIT, Cambridge, MA, October 1953:35–36.

Yngve, Victor H. 1955a. Sentence-for-sentence translation. *Mechanical Translation*, 2(2):29–37.

Yngve, Victor H. 1955b. Syntax and the problem of multiple meaning. In William N. Locke and A. Donald Booth, editors, *Machine Translation of Languages: Fourteen Essays*. Technology Press of the Massachusetts Institute of Technology and Wiley, Cambridge, MA, and New York, pages 208–226.

Yngve, Victor H. 1956. Gap analysis and syntax. *IRE Transactions on Information Theory*, IT-2(3):106–112.

Yngve, Victor H. 1957. A framework for syntactic translation. *Mechanical Translation*, 4(3):59–65.

Yngve, Victor H. 1960a. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

Yngve, Victor H. 1960b. Research objectives. *Quarterly Progress Report of the Research Laboratory of Electronics*, MIT, Cambridge, MA, October 1960:183.

Yngve, Victor H. 1961a. The COMIT system. In H. P. Edmundson, editor, *Proceedings of the National Symposium on Machine Translation*, pages 439–443, University of California, Los Angeles, February 2–5, 1960. Prentice-Hall, Englewood Cliffs, NJ.

Yngve, Victor H. 1961b. Random generation of English sentences. In *1961 International Conference on Machine Translation of Languages and Applied Language Analysis*, pages 66–80. Teddington, Middlesex.

Yngve, Victor H. 1964. Implications of mechanical translation research. *Proceedings of the American Philosophical Society*, 108(4):275–281.

Yngve, Victor H. 1967. MT at M.I.T., 1965. In A. D. Booth, editor, *Machine Translation*. North-Holland, Amsterdam, pages 451–523.

Yngve, Victor H. 1982. Our double anniversary. In *Proceedings of the*

20th Annual Meeting of the Association for Computational Linguistics, pages 92–94. Ontario.

Yngve, Victor H. 1986. *Linguistics as a Science*. Indiana University Press, Bloomington.

Yngve, Victor H. 1996. *From Grammar to Science: New Foundations for General Linguistics*. John Benjamins, Amsterdam/Philadelphia.

Yngve, Victor H. 2000. Early research at M.I.T.: In search of adequate theory. In W. John Hutchins, editor, *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*. John Benjamins, Amsterdam/Philadelphia, pages 39–72.

Yngve, Victor H., and Zdzisław Wąsik, editors. 2004. *Hard-science Linguistics*. Continuum, London/New York.