Version de travail
Paru dans :
Léon J., 2007, “ From universal languages to intermediary languages in Machine Translation : the work of the Cambridge Language Research Unit (1955-1970) » *History of Linguistics 2002* (Eduardo Guimaraes & Diana Luz Pessoa de Barros eds), Amsterdam & Philadelphia : John Benjamins Publishing Company :123-132.

From universal languages to intermediary languages in Machine Translation : the work of the Cambridge Language Research Unit (1955-1970)

Jacqueline Léon
CNRS, Université Paris 7

UMR 7597 CNRS “Histoire des théories linguistiques”
Université Paris 7
UFR de linguistique
case 7034
2, place Jussieu
75005 Paris
France
jacqueline.leon@linguist.jussieu.fr

Introduction

The Cambridge Language Research Unit (CLRU), founded in 1955 to start experiments in Machine Translation (MT), gathered many different and remarkable personalities : Margaret Masterman (1910-1986), the director of the group and a Wittgenstein’s pupil ; R.H. Richens (who died in 1984), a biologist specialist of plant genetics ; and linguists such as Martin Kay and MAK Halliday; computer scientists, among them Yorick Wilks who became one of the first researchers on Natural Language Understanding. The originality of the CLRU is that it is the only MT group, besides the Russians, to develop a method of Machine Translation using intermediary language.

The most striking aspect of their work is that the construction of this intermediary language stems directly from the 17th-century universal language schemes.

1) The universal language Nude, conceived by Richens, is widely inspired by Dalgarno’s Ars Signorum (1661) and Wilkins’ Essay (1668)[1].
2) The CLRU also makes use of the Thesaurus published in 1852 by Roget, one of Wilkins’ continuators[2].

It may seem strange that the universal language issue was considered seriously by scientists two centuries later whereas these schemes, at their time, came to almost nothing. To enlighten this point one may assume that the issue of the feasibility of MT

[1] George Dalgarno (1626-1687) ; John Wilkins (1614-1672).
[2] See Salmon (1979a) on this point.

in the 1950s raised questions comparable with those raised by philosophers in the 17th century[3] .

In my paper, I will address these issues by examining the various versions of MT methods using intermediary languages proposed by the British group in the 1950's. I will try to explain how the achievement of a practical task, the automatization of translation, as well as the implementation of a specific conception of word meaning, modified the notion of universal language itself.

1. Historical and intellectual context

Although the two periods present some analogies, it is not relevant here to compare the historical, economic and intellectual context of the apparition of universal languages in Great Britain in 17th century with the 1950s context of apparition of MT. However it is interesting to note that, in both cases, universal languages and intermediary language schemes were anchored in a strong social demand for interlingual means of communication[4].

In the 1950s, facing with the internationalization of science and the politico-military requirements of the cold war, multilingual communication technologies were greatly needed. Within a frame of unprecedented technological development as electronical computers, MT had to play a leading part in responding this demand.

The main point here is to see if the issues raised by the scientists on the feasibility of MT in the 1950s can be compared with those raised by the authors of Universal language schemes. Weaver's conceptions are very enlightening on this point[5]:

The imperfection of languages was a recurrent issue for conceptors of universal languages. It is the same for ambiguity and polysemy in MT which is one of the most complicated problems to solve by machines. Besides, in Weaver's viewpoint, the connection between translation and cryptography led " very naturally " to the idea that translation makes deep use of language invariants. Hence the translation between two languages cannot be done word by word but only by using a universal language " the real but as yet undiscovered universal language " (Weaver, 1949, p.23) ; such a project requiring considerable work on the logical structure of languages.

2. Nude: from universal language to intermediary language

[3] "Philosopher" also means "scientist" in the 17th century

[4] On the context of apparition of universal language schemes in 17th century in Britain, see Cram (1985) and Salmon (1979c, 1992).

[5] Thanks to his Memorandum " Translation " which was widely distributed among scientists in 1949, Warren Weaver (1894-1978) promoted MT in Great Britain and in the USA.

Nude is the first project of intermediary language devised by the CLRU. Although it may look rough and based upon naive conceptions of meaning and translation, it is worthy of interest because it is the first time that semantic information and primitives were used in natural language processing[6]. Besides it raised interesting questions about language representation. Like all the members of the group, Richens shared the idea of pre-eminence of semantic analysis upon syntax. This idea was very original among MT pioneers who, for most of them, thought morphology and syntax were dominant in the process of MT[7].

His experience in word by word translation led Richens to introduce semantic information to solve ambiguities.

To use semantic information in procedures, Richens proposed to build an interlingua where structural distinctive features of source languages are suppressed. Interlingua is devised as a "semantic net of naked ideas ", hence its name Nude. For Richens, semantic networks is what remains invariant during the translation process.

Nude is conceived as an algebraic language ; it comprises about fifty elements, each of them denoting a basic (naked) idea, such as plurality, plant or negation, represented by a letter.

Here are nineteen elements (out of fifty) which were used for the translation from Japanese to English of the sentence " the percentage of matured capsules and the number of grains of seeds of one capsule are different according to the time of hybridizing " :

B	becoming, change	p	plant
c	straight, plane	P	plurality, group, number
C	causation, influence	Q	hard, firm
f	possibility, potentiality	S	same, equal
H	pertain	T	time, period, duration
I	in, inside	u	elongate
L	living, alive	x	textile
M	much, more, great	X	part, component
n	near, adjacent, together	z	negation, opposite, contrary
N	contact, adhere, attach		

Nude has a syntax. A word is regarded as a relation with either 0 adjunction ; or one adjunction [.] : for example an adjective or a transitive verb expect a noun as an adjunct;

[6] There are also technical issues for the use of intermediary languages in MT. Because they constitute semantic representations common to every languages, they require far less algorithms and dictionaries than transfert methods which necessitate two algorithms for each pair of source and target languages.

[7] Note that syntax, excluding any other linguistic area, will be at the heart of Computational Linguistics, after the ALPAC disaster and the wreckage of MT in 1966.

or 2-adjunctions [:] transitive verbs[8]. Apostrophes and quotes are used as brackets within a word.

During the translation process, the source text is divided into chunks, the minimal sense units. The result of the translation of chunks into Nude is called a formula. Here is the formula of “ one seed ” :

.Pz = one   Xp’CL= seed   Xp’CL.Pz = one seed

Richens had probably read Wilkins’ Essay. As a network of semantic primitives represented by letters, Richens’ interlingua is very close of a universal characteristics[9] However Richens’ componential representation of word meaning is closer to Dalgarno’s than to Wilkins’. Remember that Wilkins’ Characteristics is devised as a hierarchical classification of concepts based on Aristotelian categories[10]. Conversely, instead of trying to codify the contents of universe as Wilkins, Dalgarno ‘s purpose was to distinguish the different semantic components of each concept, to give a sign to each component and to form the name of objects and concepts by combining the signs of all their components[11].

3. The epistemological status of intermediary language

Richens’ interlingua was widely discussed by the Cambridge group. One of the problems raised by Nude was its lack of empirical foundation in natural languages. As Wittgenstein’s pupil, Margaret Masterman could not consider Nude primitives as universal concepts. Besides she was impervious to any cognitive hypothesis according to which primitives could be the elements of a language of thought - such as Fodor’s Mentalese created a few years later (1975). For Masterman an intermediary language could not be an universal language.

She agreed with Wittgenstein that the logic unit for studying language should not be word nor proposition but word context, namely word use. In Masterman (1954, p.209) she defined use and usage in the following way : “ the Use of a word is its whole field

[8] Richens’ syntax can be viewed as of prefiguration of case grammar: a transitive verb is marked to expect a subject and an object.

[9] Moreover, just like Wilkins, he raises the issue of verbal particle compositionality (Cram, 1994).

[10] In his Essay Part 4 “A Real Character and a Philosophical Language ”, Wilkins gives the translation of the prayer *the Our Father* in fifty-one languages. In particular *bread* (p.454) is translated into his characteritics as *sɑba*, where *sɑ* denotes the genus of *Oeconomical provisions*, *b* the first difference, and *a* the second species (*bread) ; s*uch a decomposition reflects Wilkins’ hierarchical classification of concepts.

[11] This difference of opinions pulled the two philosophers apart for ever whereas they had worked together on a common scheme of universal language for a long time (Salmon, 1979b).

of meaning, its total “spread ”. Its usages, or main meanings in its most frequently found contexts, together make up its Use ”.

Because of its structure, based on the classification of words according to a set of contexts, Masterman chose thesaurus organization to create a new intermediary language, “ a thesauric interlingua ”.

For Margaret Masterman (1959, p.34) “ the fundamental hypothesis about human communication which lies behind thesaurus making is that, although the set of possible uses of words in a language is infinite, the number of primary extra-linguistic situations which we can distinguish sufficiently to talk to one another is finite. Given the complexity of the known universe it might be the case that we refer to a fresh extra-linguistic situation every time we create a new use of a word. In fact we do not ; we pile up synonyms, to rerefer, from various and differing new aspects, to the stock of basic extralinguistic situations which we already have. ”

The consequences for MT are important. Communication and translation depend on the fact that two people and two cultures, however much they differ, can share a stock of extra-linguistic contexts.

This is how Masterman defined language universality. The idea of an intermediary language refers to a stock of extra-linguistic contexts, which can be represented by a thesaurus[12].

4. Roget’s Thesaurus

Roget’s Thesaurus, in spite of its drawbacks such as incohence and non-systematicity, was chosen by the CLRU to build an interlingua combining a thesaurus with Nude.

Peter Mark Roget (1779-1869) quotes Wilkins and is considered as one of his continuators. As Wilkins, he was a philosopher and the secretary of the Royal Society. What is common between Wilkins’ Essay and the Thesaurus is the classification of words based on concepts.

However Roget took care not to build a universal language scheme. His purpose was essentially pedagogical as is indicated by the title of his book.

Roget’s Thesaurus, taking Wilkins’ Essay as a model, is divided into two parts : a thematic thesaurus and an alphabetical index. The thematic part comprises six primitive classes (abstract relations, space, matter, intellect, volition, emotion) divided themselves

[12] The CLRU was also influenced by the contextualists of the London school, namely by John Ruppert Firth (1890-1960). Firth only attended to the first meeting of the CLRU in 1955 but was not very interested in MT himself. However MAK Halliday, one of his most famous pupils, was an active member of the CRLU from 1955 till the beginning of the 1960s.

into sections then into heads. Heads are followed by a list of words connected semantically. A word can appear in several lists under different heads or classes.

To build an interlingua from Roget's thesaurus, it is necessary to have a set of coherent heads. These (arche)heads will be provided by Nude primitives. As heads can belong to several archiheads, they are classified according to a multiple hierarchy, and not only a tree organization, as in Roget's. To formalize such a thesaurus, the CRLU members chose the Lattice theory.

5. Thesauric Nude

For Masterman, the thesauric interlingua is not an algebraic universal language where elements are represented by letters. Actually the archiheads are English words. As basic semantic categories " archeheads must be below the meaning-line ". They are not words which could exist in any language. But they must be sufficiently like words which can be handled in any language . Archehead TRUE! must be like true; or at least TRUE! must be more like true than it is like please.

To Masterman the interlingua should be a genuine language, able to cope with problems of meaning such as metaphors. It is worth mentioning that at a certain point the CLRU members considered using Basic English or Esperanto as intermediary languages. Anyway what is at stake in new Nude is more the representation of natural language meaning than the universal representation of knowledge.

Here are fourteen Nude primitives (out of fifty) (Masterman 1959, p.62)

	NUDE ELEMENT	APPROXIMATING AREA OF MEANING	EXAMPLE
1	BANG !	Sudden action	Bang:think (idea)
2	DONE	Completed action	(done:change):folk (banquet)
3	WILL	Deliberate Intention	For:(will:do) (try)
4	MUCH	A lot of	Have/(much:(count: (part: where))) (long)
5	FOR	Motive,Because	For: (wil1: do) (try)
6	CAUSE	Causative actions	Cause/(have/sign) (say)
12	IN	Be Situated In, or having the Property of Being able to Contain Something	In : thing (container)
13	HAVE	Pertain " of "	Cause/nothave/life) (kill)
…	…	…	…
41	SIGN	Symbol (any sort)	Cause/(have/sign)(speak)
45	GRAIN	Pattern(artistic , thought)	Think:(stuff:grain) (chemistry)
46	HOW	Mode , quality, adjective	(think/same) : how.

47	WHEN	Time	Count:(part:when)(unit of time)
48	WHERE	Space	Change/where (move)
00	NOT	Causes all Nude elements to mean their opposites.	

Syntax is nearly the same as in Richens' version. [:] connects one element and its adjunct ; [/] is a verbal connector between subjet and verb or verb and object. For example to speak is in Nude cause/(have/sign)

Brackets replace apostrophes and quotes and unit primitives in pairs. As in Richens the process is recursive.

speak	he says	speaker
cause / (have /sign)	man/(cause / (have /sign))	man:(cause/ (have /sign))

The issues raised by thesauric Nude were meaning abstraction and category attribution. Thus the CLRU members had to find means to extract primitives from texts experimentally. From their point of view the only justification of meaning abstraction is of pratical order. They do not believe in universal knowledge representation. What is taken from universal language tradition is the empirical tradition. Just as British universal language schemes were always anchored in technological developments and social demand, such as stenography cryptography, logarithms, printing characters, language planning, multilingual communication (Cram, Maat 2000), the CRLU aimed at devising MT and information processing systems.

These practical options had great impact on the role of language formalization which was one of the main topics discussed by the group. The CRLU, boosted by MT objectives , implemented a conception of language formalization which was based on reflections upon context and meaning, independently and in competition with Bar-Hillel's[13] and Chomsky's logico-mathematics hypotheses (Léon, 2000).

For the CLRU language must be considered as a whole, and mathematically formalizable only in a second step. Whereas for Bar-Hillel it is the opposite : language is considered as mathematically formalizable a priori; and it is the task for the researcher to discover how natural languages can be adapted to formalization.

6. Wilks and templates

I will conclude this paper by mentioning the works of one of the youngest members of the CLRU, Yorik Wilks, who continued the work on Nude in the USA in the late 1960s

[13] *The Essays on and in Machine Translation by the Cambridge Language Research Unit*, where the thesaurus method was presented, were dedicated to Bar-Hillel in response to a first version (1959) of his critical record on MT (1960).

within the very new domain of Natural Language Understanding. Wilks modified Nude in order to resolve semantic ambiguities in texts. He radicalized the CLRU conception of ambiguities which should be defined with reference to dictionaries, which is the common view of MT experimenters, but within a text. He was then led to develop what he called " preferential semantics ": for a given text, a specific meaning is chosen preferably over another, so that no definitive choice should be made.

To solve ambiguities, he devised a semantic representation system for texts using "templates" which captures the "gist" of text message. Templates are pattern-matching formulas representing the meaning of a clause. These formulas are very close to Richens' but instead of encoding the various meanings of a word, they encode the meaning representation of a clause.

Here are some of the fifty-three primitives used by Wilkes to build formulas, of which forty-five are thesauric Nude's archeheads:

BE
BEAST
CAN
CAUSE
CHANGE
COUNT
FORCE
FROM
GRAIN
HAVE
HOW
IN
MAN
MAY
MORE
MUCH
MUST
ONE

Here is the formula representing the meaning of "colourless":

```
(COLOURLESS (((( ( (WHERE SPREAD)(SENSE SIGN) )NOT HAVE) KIND)
(COLOURLESS AS NOT HAVING THE PROPERTY OF COLOUR))))
```

A formula is a pair, the first part is the head COLOURLESS, the second part is a new pair which represents the translation of the word into primitives (((WHERE SPREAD)(SENSE SIGN))NOT HAVE) KIND) and its definition in natural language (COLOURLESS AS NOT HAVING THE PROPERTY OF COLOUR).

"The formula in that sense-pair can be explained as follows : 'colourless' is a sort ; a sort indicating that something does not possess some property ; the property is an abstract sensorial property of a certain sort ; that certain sort has to do with spatial extension. Thus the meaning of the whole word-sense is 'a sort that lacks an abstract, sensory, spatial property', and it is not difficult to see that this is what (in right-left order) the formula conveys." (Wilks, 1972, p.107)

Coded in LISP which was already the programming language of Artificial Intelligence, the formulas are recursive and dynamic. The first element of the list, the head, is a function, the rest of the list is an argument. Thus this meaning representation is also a computer procedure. Through pattern-matching procedure, word representations are compared with word representation in the text. If the same primitives are in the same clause, they help solving ambiguity by offering a preferred meaning.

Wilks' works had the merit of introducing semantic primitives, conceived within MT research, into the new field of non referential semantics and artificial intelligence. He chose a far more radical semantic position than Katz and Fodor (1963) since his analysis units are not grammatically correct sentences but texts. His works within the CLRU allowed him to work on semantics preferentially while syntax was dominant. This research on word and text meaning was strongly anchored in the British tradition, which was pratically based. Besides the CLRU works fit in with the 20th century British contextual tradition which was empirically based.

References

Primary sources

- Bar-Hillel Yehoshua, 1960, "The present Status of Automatic Translation of Languages" Advances in Computers vol.1, F.C. Alt ed. Academic Press, N.Y., London: 91-141.

Roget Peter Mark, 1852, Thesaurus of English words and phrases classified and arranged so as to facilitate the expression of ideas and assist in literary composition, London: Longman [2ème édition, 1853]

- Firth John Ruppert ,1957, Papers in linguistics 1934-1951, Oxford University Press
- Masterman Margaret, 1954 "Words",Proceedings of the Aristotelian Society : 209-232
- Masterman Margaret, 1959, "What is a thesaurus? " in Essays on and in Machine Translation by the Cambridge Language Research Unit. rapport non publié, ML90
- Katz Jerrold J. et Fodor Jerry A., 1963, "The structure of a semantic theory", Language vol.39, n°63: 170-210.
- Richens R.H., 1955 , "A general programme for mechanical translation between any two languages via an algebraic interlingua" [archives du CLRU, ML5]
- Weaver Warren, [1949] 1955, "Translation" in Machine Translation of Languages, 14 essays, (W.N.Locke and A.D. Booth, eds.), MIT et John Wiley **:**15-23.
- Wilkins John, 1668, An Essay towards a real character and a philosophical language London: S. Gellibrand and J. Martin
- Wilks Yorick, 1968, "On line Semantic Analysis of English Texts", Mechanical Translation, vol 11, n°3-4 :59-72
- Wilks Yorick A. 1972 Grammar Meaning and the machine analysis of language. London Routledge and Kegan Paul.

Secondary sources

- Cram David, 1985, "Universal Language Scheme in 17th century Britain", Histoire Epistémologie Langage vol 7-2: 35-44
- Cram David, 1994, " Collection and Classification : Universal Language Schemes and the Development of Seventeenth –Century Lexicography ", Anglistentag 1993 Eichstätt Proceedings, edited by Günther Blaicher and Brigitte Glaser, Tubingen : Max Niemeyer Verlag : 59-69
- Léon Jacqueline, 2000, "Traduction automatique et formalisation du langage. les tentatives du Cambridge Language Research Unit (1955-1960)", in The History of Linguistics and Grammatical Praxis (eds. P.Desmet, L.Jooken, P.Schmitter, P.Swiggers) Louvain / Paris, Peeters: 369-394.
- Maat Jaap, Cram David, 2000, " Universal Language Schemes in the 17th Century ", in History of the Language Sciences, An International Handbook on the Evolution of the study of Language from the Beginnings to the Present Edited by Sylvain Auroux .

E. F.K. Koerner Hans-Josef Niederehe Kees Versteegh, Volume 1 Walter de Gruyter . Berlin. New York :1030-1042

Salmon Vivian, 1979a, "‘‘John Wilkins’ Essay (1668): Critics and Continuators",The study of language in 17th century England, Amsterdam, John Benjamins: 97-126.

Salmon V., 1979b "The evolution of Dalgarno’s "Ars signorum’",The study of language in 17th century England, Amsterdam, John Benjamins: 157-175.

Salmon Vivian, 1979c, "Language-planning in 17th century England; its context and aims",The study of language in 17th century England, Amsterdam, John Benjamins: 129-156.

Salmon Vivian, 1992, "Caractéristiques et langues universelles", Histoire des Idées Linguistiques t.II , Mardaga, Liège :407-423.