From Weaver to the ALPAC Report

Gabriella Pardelli, Manuela Sassi, Sara Goggi

Istituto di Linguistica Computazionale, CNR, Pisa
[gabriella.pardelli,manuela.sassi, sara.goggi]@ilc.cnr.it

Abstract

This paper presents a sample pertaining to the creation and the use of words in the field of Natural Language Processing (NLP) in the years 1949-1966. These words have been statistically sorted and the results could be taken as a proof that electronic processing of linguistic data leads to the diffusion of clear and concise words for describing a complex concept which would need a circumlocution to be described instead. The aim of this article is to provide an evolutionary overview of these new lexical forms in the various languages for the period taken into account and, whereas possible, a data register and a tabular representation have been prepared as well.

1. Introduction

The study of natural language substantially developed especially after the Second World War, in particular for what concerned the application of methods of analysis. Thus language sciences rapidly evolved and one of the main elements of this evolution was surely represented by the growing inter-disciplinarity between themes and techniques proper to linguistics and the ones pertaining to other sciences: just to give an example, the use of logical-mathematical methods and of statistical and quantitative ones for studying languages and literary works.

Historically, the interest in researches in the field of automatic translation was stimulated by a famous memorandum spread by Warren Weaver in 1949; in the '50s Father Roberto Busa started his researches on the automatic compilation of concordances in the complete works of Saint Thomas d'Aquino, thus enriching the field of lexicographical studies (still alive today) with his experience and his own methods.

The use of personal computers in linguistic researches and analyses was finally consacrated at various international conferences (Strasbourg, 1957; Besançon, 1961; Nancy, 1964; Pisa, 1968).

In the '50s and '60s, adjectives such as *automatic* and *mechanic* and nouns such as *mechanisation* and *machine* were often used for describing computational approaches in the field of natural language studies.

The following are some selected titles published in the period:

Papers on Mechanical Translation (Bar-Hillel et al., 1952); *Mechanical Resolution of Linguistic Problems* (Booth et al., 1958); *La traduction automatique* (Panov et al. '58); *La Machine à Traduire*, (Delavenay, 59); *La Mécanisation dans les Recherches Lexicologiques* (Quemada, 1959); *Linguistic Analysis and Programming for Mechanical Translation* (Ceccato, 1960)[1]; *Les travaux lexicologiques préparatoires à la traduction automatique* (Pottier, 1961); *Automatic Translation of Languages* (Ghizzetti, 1962); *L'automation des recherchers documentaires. Un modèle générale: Le Synto*l (Cros, Gardin et Lévy, 1964); *Readings in Automatic Language Processing* (Hays, 1966); *Les Machines dans la Linguistique* (Stindlová, 1966); *Traitement de l'information linguistique par l'homme, par la machine* (Deweze, 1966); *Machine Translation* (Booth, 1967); *Automated Language Processing* (Borko, 1967); *2ème Conference Internationale sur le Traitement Automatique des Langues* (CITAL, 1967); *La critique des texts et son automatisation* (Froger, 1968).

In these years, the use of a computer in linguistics and therefore for linguistic data (graphemes, words, sentences, texts, etc.) processing, was described by means of adjectives such as *mechanic* and *automatic*. Along with these adjectives, the nouns *mechanization* and *automation* were applied for globally representing the computational functions in this field, such as, for example, information retrieval from literary texts.

In 1966, the term "*Computational Linguistics"* was used for the first time in an official document, the famous ALPAC Report of the *Automatic Language Processing Advisory Committee* on Computational Linguistics and Automatic Language Translation. Till then, the terms *Automatic Language Processing* and *Mechanolinguistics* had been used with nearly the same meaning.

The noun *Computer* was used since the '60s for indicating automatic systems for linguistic researches and analyses. Here are some examples: *Computer Applications in the Behavioral Sciences* (Borko, 1962); *Natural Language and the Computer* (Garvin, 1963); *Mechanization and Automation in Linguistics. Application of Elecronic Computers* (Sgall, 1964)[2]; *A Computer-aided Study of Literary Influence: Milton to Shelley* (Raben, IBM-64).

In the 1960s, adjectives such as *automatic* and *mechanic* and the noun *computer* cohexisted until the adjective *computational* was introduced: it was going to be widely used by experts of the field by the end of the '60s-beginning of the '70s (for example, *Computational Analysis of Present-day America English* (Kučera & Francis, 1967).

[1] Technical Report no. RADC -TR-60-18 prepared for European Office; Air Research and Development Command, Air Force, U.S., by Centro di Cibernetica e di Attività Linguistiche, University of Milan. Milano, G. Feltrinelli Editore, [1961].

[2] In P. Sgall (ed.). *Cesty moderní jazykovedy*, pp. 150-159, 1964, Praha, Orbis.

2. Methodology

2.1 How to measure information

The process of managing data, combining retrieved information in order to define a significant and meaningful sample, it is not always an easy task and methodological doubts arose especially at the beginning.
The first step of this work has been to determine and then gather data to be measured, that is NLP terminology which developed and spread in the years 1949-1966.
The problem of an appropriate information retrieval underlies the process of sampling, since single units should first be measured and then gathered for creating the sample: which is the best procedure for storing data? how should information be measured?
The criterion adopted for retrieving data does not pertain to type of analysis (i.e. text processing, automatic translation, etc.) or to the type of publication (i.e. Conference Proceedings, journals articles, monographies, etc.): it has been decided to focus on the names of those authors who dealt with automatic language processing of linguistic data. The result of a semantic analysis of data would have been a fragmentation of the sample, then partial samplings on available data have been made: these samplings were not always very satisfactory and, especially in the final phases of the process, manual interventions have been necessary in order to retrieve more information from library catalogues.

2.2 Information retrieval

Words from publication titles have been retrieved by means of searches on web sites of major libraries. At first, the attention has focused on selecting bibliographical entries, for eliminate scarcely significant or non-pertinent data and concentrating only on pertinent information. The bibliographical fields of the 8,256 selected records have then been reduced to only three: title, year and name of the author. The collection of data has started at the beginning of 2003, the textual processing and indexing has been done at the end of 2003.
In the first phase of data storing, an heuristic approach has been adopted but afterwards it has been necessary to choose criteria for data retrieval. The searches on the web sites of major data bases such as National Libraries of several countries by keywords and years of reference (1949-1966) have not been very satisfactory, because data extracted from words contained in titles could refer to numerous disciplines. The manual correction of non-pertinent or scarcely significant entries would have taken too much time and it was therefore necessary to define a system of queries in order to obtain reliable and suitable results for creating the sample: information retrieval by means of significant words showed to be the more reliable.
In the second phase, the information gathered from web sites of many libraries has been tested and then only three libraries have been selected: the British Library, the Library of Congress and the Bibliothèque Nationale de France[3].

[3] http://www.loc.gov/, http://www.bl.uk/, http://www.bnf.fr/

The following are just a few examples: the result of the search of the word *machine* on the Subject Search menu of the British Library is 40,895 records, while searching the same word on the Author/Title Search menu records are 30,874. There are 11,233 records for the word *computational* sorted by the Author/Title menu.
The same search on the Library of Congress data base gave these results: 1,850 records for the word *machine* and 969 records for the word *computational*, both sorted by title.
The combination of keywords and a careful analysis of the results, provided the first amount of information for creating the archive. Nevertheless, data was not enough yet: information has then also been retrieved from the telematics net, in particular from titles of presentations at international conferences in the field until 2003; other useful information has been gathered from the web site of the Association for Computational Linguistics (ACL); and lastly, from a selection of texts indexes available at the library of the Institute of Computational Linguistics, especially those of the *Donazione Zampolli* (for example, ALPAC-66; CITAL-67; IBM-64; ICCL-69; Weaver, 1949). For these texts, the analysis have been carried out on bibliographical references.
These analyses of bibliographies of the reference period have provided further data with queries on the name of authors (i.e. Bar-Hillel, Y.; Booth, A.D.; Burton, D.; Delatte, L.; Dubois, C.; Evrard, E.; Garvin, P.L.; Harris, Z.S.; Kay, M.; Minsky, M.L.; Oettinger, A.G.; Panov D.I.; Schank, R.C.; Sebeok, T. A.; Tollenaere, F. de; Wood, W.A.; Vauquois, B.; Ziff, P.) and on precise keywords (i.e. analyse automatique; automatic translation; Computer and the Humanities; language and machines; mechanical translation; traduction automatique).
The phase of reading of articles and selection of bibliographies has been very time-consuming.

2.3 Analysis of data

The second part of the work has mainly consisted in the analysis of retrieved data (the keywords extracted from bibliography titles) once the available amount has been considered sufficient for creating the sample, starting from the years of reference (1949-1966) and making a comparison until 2003. Data have been processed with DBT software (by Eugenio Picchi, CNR patent) and then a textual archive has been created: by consulting the database and applying DBT functions, it has been possible to extract from the archive the contexts relating to the chosen keywords.

automata	automate	automated
automatic	automática	automatically
automatico	automating	automation
automatique	automatiquement	automatiques
automatisation	automatische	automatischen
automatisee	automatisierte	automatisierung
automatism	automatized	automaton
automatycznym	computatio	computability
computacionales	computation	computationaily
computational	computationally	computations
compute	computed	computer
computerdialogue	computergestützten	computerized
computerizzata	computerlexicon	computerlinguïstiek
computerlinguistik	computerphilologie	computers

computes	computing	machina
machine	machinery	machines
mechanical	mechanism	mechanisms
mechanization	mechanized.	

3. Graphical representation of the sample

The last phase of synthesis consisted in a graphical representation of words retrieved with the string search *autom-*. The value taken into account for representing the results is distribution of words over the years (see Fig. 1).

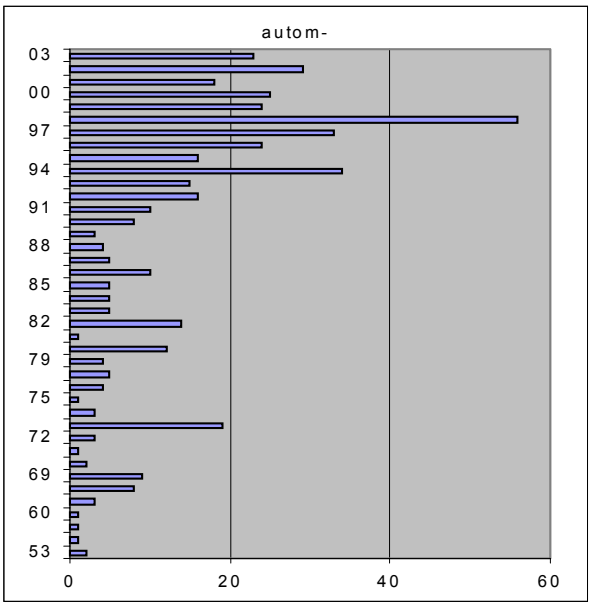

Figure 1: The string search *autom-*

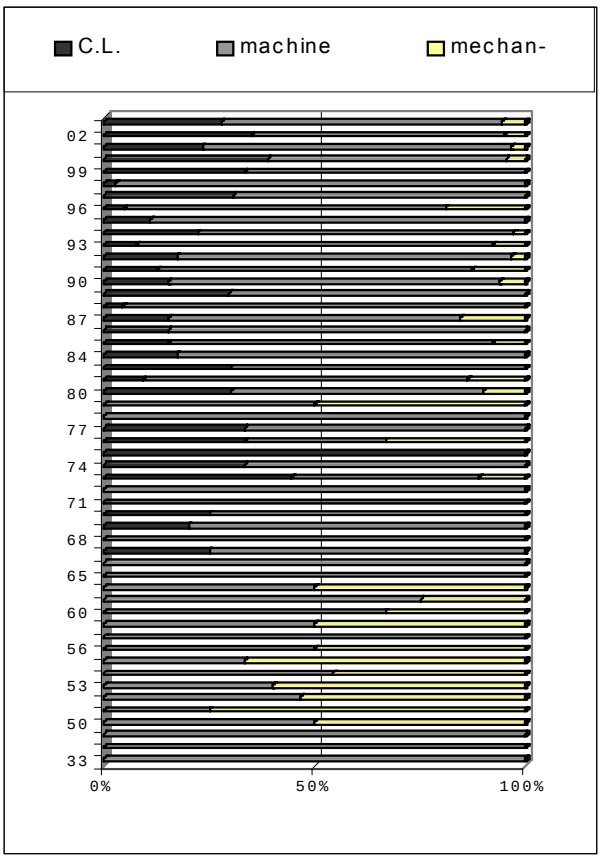

Figure 2: Terms Comparison

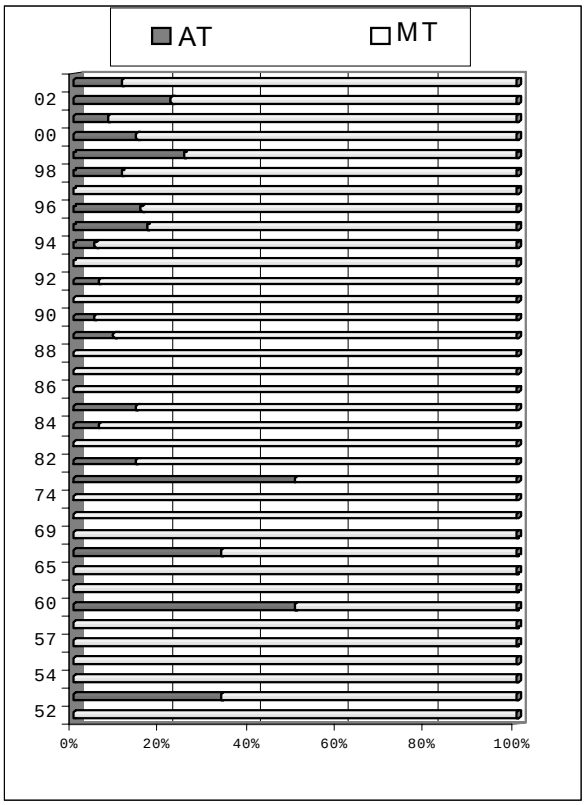

Figure 3: Use of MT and AT

Figure 2 shows the use of the following terms over the years: Computational Linguistics (CL), the search string *mechan-* and *machine*. Since 1979 the trend has been to use the term *mechanism* instead of mechanical and similar. Figure 3 shows a comparison between *Machine Translation* (MT) and A*utomatic Translation* (AT).

4. Conclusions

From the end of the 1940s since the beginning of the 1960s, processing of linguistic data was mainly focused on various attempts of automatic translation and on text processing; linguists started to accept the use of computers in their field, debating and making plans for the future at major international conferences. In this decade a common term for indicating the new directions of the research in the field did not emerge: they could have been classified as researches of applied linguistics and mathematical linguistics as well. In this case, such definitions have not been taken into account in this survey: terms like *Applied Linguistics* and *Mathematical Linguistics* would have largely broadened the analysis to the detriment of the most significant ones.

The use of the computer in linguistic researches never stopped since then: from the lexicographical initiatives of the European research centres, to the great projects of automatic/machine translation in the United States, for finally arriving to the stage of using the *machine* in applications of linguistic engineering. This means that linguistic resources have become essential for a deeper knowledge of languages.

Philological disciplines have benefited from these applications: for example, textual processing of Hippocrates' works (*Concordantia in Corpus*

Hippocraticum 1986) have been fundamental for attributing the paternity of works considered as spurious.
In the last 50 years the growing inter-disciplinarity due to the use of computers, has favoured the development of new perspectives and the creation of resources in many fields. If the ALPAC report is considered as the milestone for the birth of Computational Linguistics directly from automatic/machine translation, Figure 4 shows the yearly growth of researches in this field since 1945.

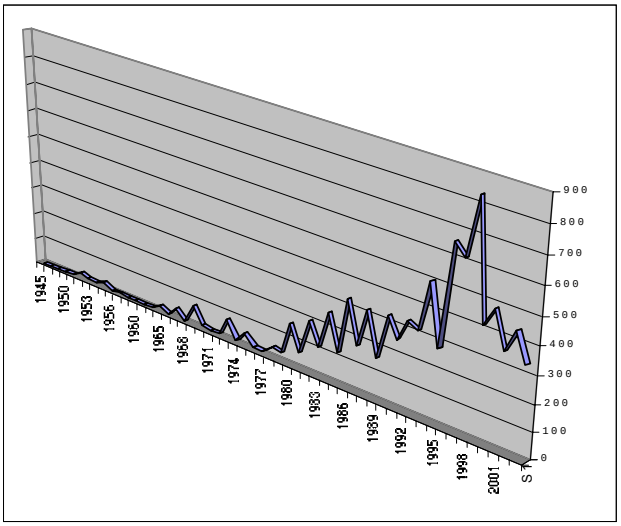

Figure 4: Diachronical representation of the archive

Nowadays researches in Computational Linguistics have so many fields of application that it is easy to found them already outlined in the recommendations of the ALPAC report, quoted here below from page 29, "Automatic Language Processing and Computational Linguistics":

1. to teach foreign languages more effectively;
2. to teach about the nature of language more effectively;
3. to use natural language more effectively in instruction and communication;
4. to enable us to engineer artificial language for special purposes (e.g., pilot-to-control tower languages);
5. to enable us to make meaningful psychological experiments in language use and in human communication and thought (unless we know what language is we do not know what we must explain);
6. to use machines as aids in translation and in information retrieval.

5. References

Actes du Colloque International sur la Méchanisation des Recherches Lexicologiques. Besançon, Juin (1961). *Cahiers de Lexicologie*, Vol. 3, 61, Publiés par B. Quemada.

Actes du Seminaire International sur le Dictionnaire Latin de Machine. (1968). Rédigé par Roberto Busa S.J. *Calcolo*, Supplemento No. 2, Vol. V (1968).

Université de Nancy, Faculté des Lettres et des Sciences Humaines (1966). *Actes du Premier Colloque International de Linguistique Appliquée*, Nancy, 1964.

Association for Computational Linguistics (1983). *First Conference of the European Chapter of the Association for Computational Linguistics*. Proceedings of the Conference, 1-2 September 1983, Pisa, Italy.

Bessinger Jess B. Jr., Parrish Stephen M., Arader H.F. (eds.) (1965). *Literary Data Processing Conference*, New York, Sept. 9-11 - 1964. IBM, Data Processing Division.

Busa, Roberto S.J, International Business Machines Corporation (1951). *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum*: A first example of word index automatically compiled and printed by IBM. Milano, Fratelli Bocca.

CITAL (1967). *2ème Conference Internationale sur le Traitement Automatique des Langues*. Grenoble.

Delavenay, Émile (1959). *La machine à traduire*. Paris, Presses Universitaire de France.

Garvin, Paul L., Spolsky Bernard (1966). *Computation in Linguistics: A Case Book.* Bloomington, Indiana University Press.

Godfrey J.J., Zampolli A. (1997). Language Resources. In A. Zampolli, G.B. Varile (Managing Editors), *Survey of the State of the Art in Human Language Technology,* Linguistica Computazionale, XII-XIII. Pisa, Giardini Editori (also Cambridge University Press), pp.381-384.

Hays David G. (ed.) (1966). *Readings in Automatic Language Processing*. New York, American Elsevier.

Hays David G. (1967). *Introduction to Computational Linguistics*. New York, American Elsevier.

Juilland Alphonse, Roceric Alexandra (1972). Analytic bibliography, in *The linguistic concept of word.* The Hague, Paris, Mouton. 11-59.

National Academy of Sciences, National Research Council (1966). *Language and machines: computers in translation and linguistics*. A Report by the Automatic Language Processing Advisory Committee, WA.

Pardelli G., Orsolini P., Sassi M., Enea A., Gazzetti S. (eds). (2002). *TAL Bibliography (1951-2002).* Pisa, S.T.A.R.

Pardelli G., Sassi M. (2001). *ILC Library: Cataloghi e Indici.* Pisa, S.T.A.R.

Pardelli G. (2003). BIBLOS: historical, philosophical and philological digital library of the Italian National Research Council, *Linguistica Computazionale*, XVIII-XIX, Vol. II, Pisa-Roma, I.E.P.I., pp.519-546.

Quemada B. (1957). La technique des inventaires mécanographiques. In *Lexicologie et Lexicographie Françaises et Romanes. Orientations et Exigences Actuelles*. Actes du Colloque International, Strasbourg. Paris, CNRS, pp. 53-63.

Stindlová Jitka, Mater Erich (Rédaction). (1968). *Les machines dans la linguistique: colloque international sur la mécanisation et l'automation des recherches linguistiques*, Praque. Academie Techecoslovaque des Sciences et l'Academie Allemande des Sciences.

Weaver, W. (1949): 'Translation'. Repr. in: Locke, W.N. and Booth, A.D. (eds.) *Machine translation of languages: fourteen essays*, Cambridge, Mass.: Technology Press of the MIT (1955), pp. 15-23.

Zampolli A. (1973). Humanities Computing in Italy. *Computers and the Humanities*, VII, n.6. New York, SED Publications, pp. 343-360.

Zampolli A., Calzolari N. (eds.) (1977). *Computational and Mathematical Linguistics*, Proceedings of the International Conference on Computational Linguistics. Biblioteca dell'Archivium Romanicum", Serie II: Linguistica, Leo S. Olschki Editore, Firenze. Vol. 36.