Natural Language Processing: A Terminological And Statistical Approach

Gabriella Pardelli, Manuela Sassi, Sara Goggi, Paola Orsolini

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

Abstract

The aim of this article is to provide a statistical representation of significant terms used in the field of Natural Language Processing from the 1960s till nowadays, in order to draft a survey on the most significant research trends in that period. By retrieving these keywords it should be possible to highlight the ebb and flow of some thematic topics. The NLP terminological sample derives from a database created for this purpose using the DBT software (Textual Data Base, ILC patent).

Scientific presentations at the main conferences of the '60s point out a frequent recurrence of expressions such as *mécanisation des études lexicologique, les machines à cartes perforées et leurs application lexicologique*, which trace back to the origin of electronic processing of linguistic data and to some solutions of linguistic-literary problems, to lexicographic researches, to scientific terminology, to automatic dictionaries, to homographs, synonyms and the possibility of producing indexes and concordances by means of an electronic processor. Terms such as *meccanizzazione, mechanical translation, machine à traduire* used by experts of the field in the 1950s and 1960s seem to well testify the change, the shift, the beginning and then the final consecration of a rapidly evolving field: Natural Language Processing.

1. Introduction

The terms introduced in the first international conferences in the field, those which witnessed the early use of machines in linguistic researches and analyses, are briefly illustrated. Among the conferences, the following have been chosen: Strasbourg 1957; Tübingen 1960; Besançon 1961; Strasbourg 1964; New York 1964; Praga 1966; Grenoble 1967; New York 1964; Pisa 1968. The choice of these specific conferences wants to underline the importance of computer applications to linguistic and literary analysis: the conferences in Strasbourg, Tübingen and Besançon witness the shift from mechanical processes to electronic processing; the 1964 Strasbourg one is relevant for quantitative linguistics; those in Praga and Pisa for lexicography; the one in New York is important for literary studies while the Grenoble one plays an important role for the wide variety of Computational Linguistics issues dealt with. Scientific presentations at the above-mentioned conferences point out a frequent recurrence of expressions such as *mécanisation des études lexicologique*, *les machines à cartes perforées et leurs application lexicologique* which trace back to the origin of electronic processing of linguistic data and to some solutions of linguistic-literary problems, to lexicographic researches, to scientific terminology, to automatic dictionaries, to homographs, synonyms and the possibility of producing indexes and concordances by means of an electronic processor: "L'utilisation des machines a introduit dans notre ordre des recherches un bouleversement total. Des travaux des type artisanal, nous passons brutalement aux technique industrielles ..." (Quemada, Besançon, 1961, p.18). A few years before, the expression *Mechanical Resolution of Linguistic problem* appeared (Booth et al., 1958). Terms such as *meccanizzazione, mechanical translation, machine à traduire* used by experts of the field in the 1950s and 1960s seem to well testify the change, the shift, the beginning and then the final consecration of a rapidly evolving field: Natural Language Processing (NLP). It is therefore the mechanical or electronic device called "machine" which will be used by NLP pioneers in those years; the noun "computer" will make its appearance later, while we will have to wait until the late '60s for the use of the adjective "computational". Data extracted from the *Computers and the Humanities* journal and from the major international conferences selected for this survey are presented, ie. Association for Computational Linguistics (ACL), in particular the ACL Anthology: titles of presentations at the various ACL-related conference have been extracted; European Association for Language Resources and Evaluation (ELRA): titles of presentations at the LREC conferences and workshops have been considered (see Figure 1).

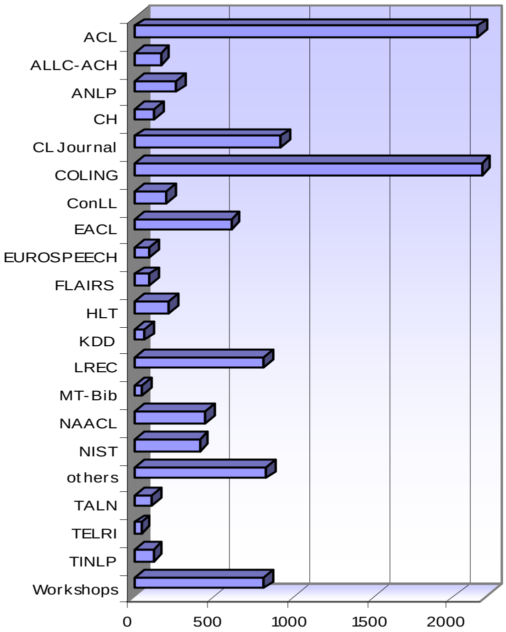

Figure 1: Composition of the Corpus.

Being such a vast field, the authors have selected and limited the data for producing the sample only to the conferences of international Associations, going back in time from the oldest to the most recent ones and thus giving the picture of a resource in constant evolution.

The NLP terminological sample derives from a database - created for this purpose using the DBT software (Textual

Data Base, CNR-ILC patent) – containing 4000 titles presented at international conferences and implemented by 10.000 titles of the field (the outcome of a previous work) and by data coming from the bibliographical analysis of the early ‘60s issues of the *Computers and the Humanities* journal.

The DBT tool has allowed to organise the material in lists sorted by title, author and year and indexes of terms in various languages; data has then been used for various kinds of statistical searches.

This work is supplied with some graphical representations of the data showing a diachronic interpretation of the statistical sample of NLP terminology.

2. A Brief History of NLP Terms

NLP can be defined as the entire set of methods and techniques for processing a language by means of a machine.

Before the computer age, the manual method for textual processing consisted in the transcription on cards of all single words with their relevant contexts, that is with the sentence they belong to. Then, cards were manually ordered in alphabetical order. Obviously, considering that - on average - for each single word a sentence made up of seven words is quoted, the entire work meant transcribing the whole text seven times.

The most relevant change which followed was the possibility of automatically managing the data; the transcription process of every single word of a text was made on special cards called: *schede meccanografiche / cartes mécanograpiques / punched card machines*. What appeared on these cards was the traditional information of a lexicographic card (lemma, author, work, reference, date) and a context of concordances which could be ‘long’ (or ‘macro-context’) in opposition to the ‘short’ (or ‘micro-context’) one. The essential data were perforated, in order to perform an automatic information extraction for the subsequent creation of indexes. The philological analysis of texts was still entrusted to philologists: on this respect, it is very significant what Louis Delatte stated: “Les machines ne nous mettent pas encore à l’abri des erreurs philologiques. Ce serait trop beau! Mais elles nous épargnent un temps considérable que, sans elles, l’on consacrerait à des besognes purement matérielles. Elles permettent en outre par des séries de tris assez complexes de vérifier si une analyse a été correcte”.[1]

The first approach of electronics to linguistics, or rather the applications of computer to linguistic analysis, dates back to the 1950s. Electronic processing of linguistic data inspired the first pioneers and the field of Natural Language Processing rapidly matured, thus providing classical lexicologists with new and unforeseen tools for varied linguistic analysis (literary, stylistic, metrical and quantitative).

Gradually, several projects came to life in research centres funded on purpose and their variety and dimension grew yearly. Most important was the start of projects on historical dictionaries which foresaw textual processing: it was then justified to talk about “the death of the hand-made concordance” (Raben, 1969)[2]. The computer was then used in projects pertaining to the creation of lexicons, for studies on terminology through textual processing of the complete works of philosophers and man of letters. In statistical linguistics the computer was immediately used for describing the quantitative features of a text or group of texts; measuring these kind of features was indeed a great help for defining the style of an author and solving problems of various nature (philological, chronological and of attribution). Quantitative studies on languages led to the production of dictionaries for the various languages at the beginning of the ‘70s (see Figure 2).

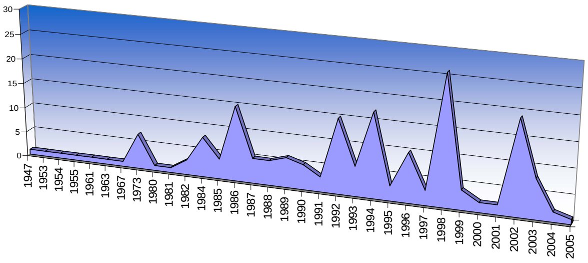

Figure 2: Occurrence of “Dictionary”

In the following paragraphs some of the terms most frequently used in those years will be analysed.

2.1. NLP: Historical Terms

Augusto Guzzo’s words[3] remind us which is the linguistic use of the word “macchina”: “*Macchina è mechané*. Quando l’intreccio d’una vicenda tragica non si scioglieva naturalmente, Euripide faceva scendere su la scena una divinità mediante un congegno di teatro: deus ex machina, theòs apò mechanês. Ma questo è un significato derivato. *Mechané* è anzitutto la trovata mentale che inventa un congegno o qualsiasi altro mezzo per sottrarsi alle difficoltà della vita”.

It is therefore the *mechanical* or *electronic* device called *machine* which the NLP pioneers started using in the ‘50s and ‘60s. The noun “machine” was normally associated with words pertinent with all the computational approaches of the field, as for example ‘*machine translation*’ (MT): the term stays for a computational system of translation with no human aid.

Consequently also the adjective *mechanical* started to be used, from the Latin ‘mechanicu(m)’ and the Greek ‘mekhanikós’ (‘mekhane’=“machine”); in fact, the expression *mechanical translation* was used by the pioneers of automatic translation (Bar-Hillel, 1952). The use of machines or data processing systems for processing linguistic data lead to the noun *mechanization* in Italian

[1] In À propos d’une concordance, *Les machines dans la linguistique*, p.192.

[2] In *Scholarly Publishing*, vol. 1, no. 1 (1969), p. 61.

[3] A. Guzzo, L’uomo, la macchina, la tecnica. Relazione introduttiva. In *L’uomo e la Macchina*: Atti del XXI Congresso Nazionale di Filosofia, I. Relazioni, Edizioni di Filosofia, Torino, 1967, p. 4.

(Ceccato, 1951), English (Westphal, 1975) and French (Rand, 1960) (see Figure 3).

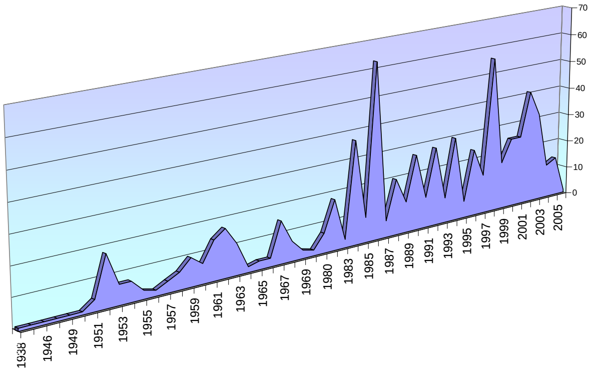

Figure 3: Occurrence of “Machine” and derived words

The tool, the device, the machine which automatically performs certain functions without any human aid or intervention suggests as well the adjective *automatic.*
The adjective *mechanical* prevailed undisturbed in the NLP vocabulary until the mid of the ‘60s, when it was outclassed by another adjective, *computational*, of course related to the first computers (see Figure 4).

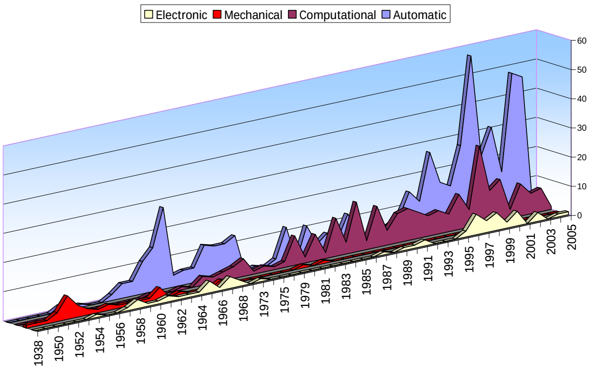

Figure 4: Comparison

The following words of Antonio Zampolli are helpful for defining terms such as Natural Language Processing and Computational Linguistics: “… I should like to make a few observations concerning its [CL] content in relation to recent developments in the area of automated language processing (ALP). I use this term rather than *computational linguistics* as it is far more general in its implications, encompassing all studies, theoretical or applied, on the use of computers or computational techniques in the processing of natural language. I consider computational linguistics to be a subset of ALP, a subset to which I shall refer by the abbreviation CL”.
The expression *automated language processing* (ALP) has been the forerunner of the expression NLP and was commonly used in the whole ‘60s (see Figure 5). In the past, terms such as computational linguistics, mathematical linguistics, algebraic linguistics, have been often been used without distinction, as they were interchangeable: in the past, it happened more often to talk about computational linguistics (CL) in academic contexts, but in reality it is rather a synonym of NLP and of nowadays Language Engineering (LE)[4].

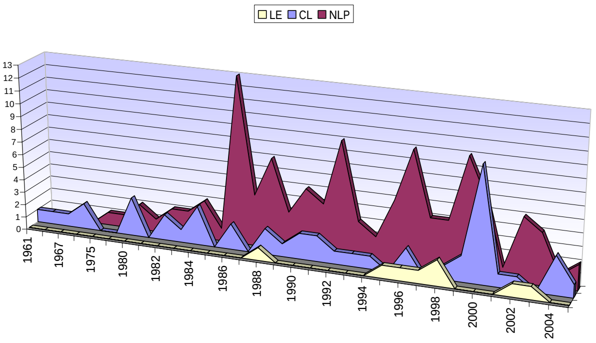

Figure 3: Comparison

3. Topics

A special attention is paid to topics, which have been analysed also separately and represent a subset of 1001 items. As a first result, here are the most frequent words:

128	language	26	applications	16	Lexical
115	speech	26	corpora	15	Extraction
67	processing	26	dialogue	15	Generation
47	evaluation	26	issues	15	Modelling
43	information	25	text	14	Acquisition
41	resources	24	natural	14	Methods
40	systems	23	analysis	14	Syntax
40	tools	21	computational	13	lexicography
38	recognition	21	corpus	13	Linguistics
35	spoken	21	multimodal	13	Semantic
34	parsing	19	multilingual	12	Data
34	translation	18	learning	12	lexicon
31	semantics	17	knowledge	12	linguistic
28	machine	16	annotation	12	mt
28	retrieval	16	discourse	12	terminology

Then couples of the most frequent words have been extracted and results are represented in Figure 5.
If the terms appearing in the table above are read as keywords, they show the main research themes of the field and among them those related to the *speech* area particularly emerge. Words such as *retrieval* and *extraction* underline the importance of information retrieval especially in the web era, while the term *multimodal* belongs to a very narrow domain of application which is nowadays most studied: Language Engineering.

[4] “To summarise: CL is a part of the science of language that uses computers as investigative tools; NLP is part of the science of computation whose subject matter is computer systems that process human language”. The quotation is from H. Cunningham, 1999 *A Definition and short history of Language Engineering*, Language Engineering, vol. 5, (1), p. 4. “Language Engineering is the discipline or act of engineering software systems that perform tasks involving processing human language”. *Ibidem* p. 5.

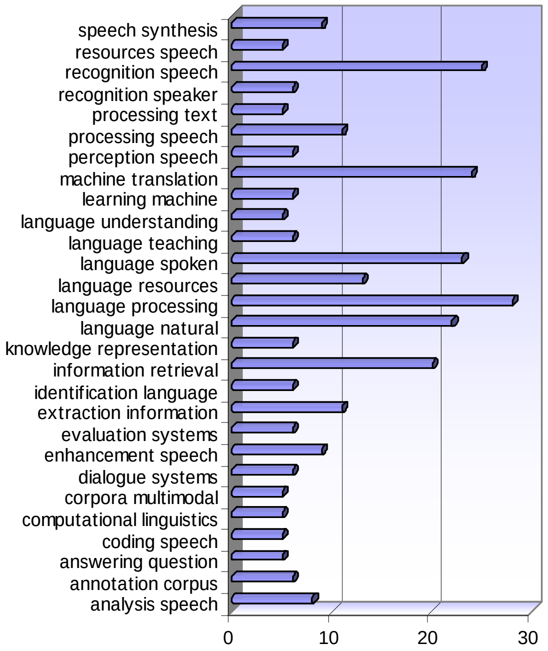

Figure 5: Statistical co-occurences

4. Conclusions

The results described in this article show how at dawn NLP used to borrow its terminology from other fields (such as computer science) and to widely adopt adjectives like electronic, automatic, mechanical and cybernetic. In time, the field developed its own terms (i.e NLP and Computational Linguistics) which represent it still nowadays.

All said, the role of terminology in today multilingual society, ever more depending on knowledge and information, it is crystal clear.

5. References

Actes du Colloque International sur la Méchanisation des Recherches Lexicologiques. Besançon (1961). *Cahiers de Lexicologie*, Volume 3.

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

Université de Nancy, Faculté des Lettres et des Sciences Humaines (1966). *Actes du Premier Colloque International de Linguistique Appliquée*, Nancy, 26-31 Octobre 1964.

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

Bar-Hillel, Y., Perry, J.W., Reifler, E., *et al.*, (1952) *Papers on Mechanical Translation*.

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

Busa, R.S.J. (1951). *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum*: A first example of word index automatically compiled and printed by IBM punched card machines. Milano, Fratelli Bocca.

Ceccato, S., *La Meccanizzazione delle Attività* Umane Superiori, parte I, in "Civiltà delle Macchine", IX, 4, Torino, 1961, pp. 22-9.

CITAL (1967). *2ème Conference internationale sur le traitement automatique des langues*. Grenoble.

Cori, M., David, S., Léon J. (2002). Pour un travail épistémologique sur le TAL. TAL Volume 43 n° 3/2002, pp. 7-20.

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

Garvin, P.L. (1972). *On Linguistic Method*. The Hague, Mouton.

Garvin, P.L., Spolsky, B. (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, *Survey of the State of the Art in Human Language Technology*, Pisa, Giardini Editori. (Also Cambridge University Press).

Guzzo, A., L'uomo, la macchina, la tecnica. Relazione introduttiva. In *L'uomo e la Macchina*: Atti del XXI Congresso Nazionale di Filosofia, I. Relazioni, Edizioni di Filosofia, Torino, 1967, p. 4.

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

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

Juilland, A., Roceric, A. (1972). Analytic Bibliography, in *The Linguistic Concept of Word.* Paris, Mouton. 11-59.

National Academy of Sciences, National Research Council (1966). *Language and Machines: Computers in Translation and Linguistics*. Washington, D.C.

Pardelli, G., Orsolini, P., Sassi, M., Enea, A., Gazzetti, S. (a cura di) (2002). *TAL Bibliography (1951-2002). Parte I.* Pisa, S.T.A.R.

Pardelli, G., Sassi, M., Goggi, S. (2004). *From Weaver to the ALPAC Report.* LREC 2004: Proceedings, Vol. VI, Paris, (ELRA). 2005-2008.

Quemada, B. (1957). La technique des inventaires mécanographiques. In *Lexicologie et Lexicographie Françaises et Romanes.* Paris, CNRS. 53-63.

Stindlová J., Mater E. (Rédaction) (1968). *Les machines dans la linguistique.* Academia, Editions de l'Académie Tchécoslovaque des Sciences. Prague.

University of Michigan. (1960). Foreign language project "Mechanization": a project for research, controlled experimentation, design and testing in the mechanization of the language learning processes.

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). 15-23.

Zampolli, A. (1973). Humanities Computing in Italy. *Computers and the Humanities*, Volume VII (6), pp. 343-360.

Zampolli, A., Calzolari, N. (1977) (eds.). *Computational and Mathematical Linguistics*. Leo S. Olschki Editore, Firenze. Volume 36.

http://www.loc.gov/

http://www.bl.uk/

http://www.bnf.fr/

http://acl.ldc.upenn.edu/