History of Computing

Jacqueline Léon

Automating Linguistics

Springer

History of Computing

The *History of Computing* series publishes high-quality books which address the history of computing, with an emphasis on the 'externalist' view of this history, more accessible to a wider audience. The series examines content and history from four main quadrants: the history of relevant technologies, the history of the core science, the history of relevant business and economic developments, and the history of computing as it pertains to social history and societal developments.

Titles can span a variety of product types, including but not exclusively, themed volumes, biographies, 'profile' books (with brief biographies of a number of key people), expansions of workshop proceedings, general readers, scholarly expositions, titles used as ancillary textbooks, revivals and new editions of previous worthy titles.

These books will appeal, varyingly, to academics and students in computer science, history, mathematics, business and technology studies. Some titles will also directly appeal to professionals and practitioners of different backgrounds.

More information about this series at http://www.springer.com/series/8442

Jacqueline Léon

Automating Linguistics

Springer

Jacqueline Léon
Laboratoire d'Histoire des Théories
Linguistiques UMR CNRS 7597
Université de Paris
Paris, France

ISSN 2190-6831 ISSN 2190-684X (electronic)
History of Computing
ISBN 978-3-030-70641-8 ISBN 978-3-030-70642-5 (eBook)
https://doi.org/10.1007/978-3-030-70642-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

With the latest comeback of artificial intelligence as 'deep learning', newspapers and industry have been promising us new computer breakthroughs in linguistics and information retrieval. Building upon automatic processing of data that are massively harvested from the web, we are promised in some not too distant future near-perfect automatic translations, expert chatbots that understand and answer human queries, or even conversations with virtual assistants. However, as some reports have brought to light, much human effort, often under precarious circumstances, is tacitly injected in the machine's 'deep learning'[1] and we do not yet know how the learning curve of this enhanced technology may evolve.

But it is not the first time that we have been promised perfect translation or improved human-computer interaction, rather, as history teaches us, the industry's self-advertising through the projection of a futuristic utopia is a recurrent phenomenon of our computer age. Already in the 1950s, formalisations of language were proposed that would supposedly make automatic translation possible. They turned out to be performing poorly. And already in the 1960s we saw the first usage of computing facilities for corpus linguistics prefiguring later big data or digital humanities. But, tempered by the limited resources of the time, this was without the (over)ambitious high hopes pinned on today's big data. This goes to show that it is now more timely than ever to go back in time and reflect upon past developments in computer linguistics. Both the successes and the limits of earlier efforts can help to historically inform us and to critically assess our current situation.

The present book is a history of how the digital computer encountered the field of linguistics in the wake of the Second World War and slowly but lastingly changed the very field of linguistics, creating new (sub)fields such as Automatic Translation, Natural Language Processing or Computational Linguistics. Two important turns are described in this book. The first one, which may be called the 'automatic turn', is the automation of language, enabled through the formalisation and

[1] Julia Carrie Wong, 'A white-collar sweatshop: Google Assistant contractors allege wage theft', *Guardian*, May 25, 2019.

mathematisation of language that took place roughly between 1949 and 1966. The second turn, the 'corpus turn', is the emergence of natural language processing in the 1990s, continuing and enlarging earlier research in documentation systems and corpus linguistics with the help of microcomputers.

Though efforts to formalise language and automate linguistics antedate this fateful encounter, the advent of the digital computer accelerated and heavily influenced the automation of language. It enabled, both theoretically and practically (and also financially), the use of mathematical methods in language, and, later, the systematic and automatic exploitation of large corpora in linguistics. But this encounter was also a two-way process. Linguistics also contributed to the newly developing field of computing. It motivated the development of some early programming languages, documentation systems and query languages, and, most conspicuously, provided some of the important theoretical tools for computing and programming such as indexing and parsing algorithms or the Chomsky hierarchy.

The main trigger for the encounter between linguistics and computing was the Second World War. Linguistics took part in the war effort as much as the other sciences, a fact that is often overlooked (cf. Chap. 3). Apart from the more obvious connection to cryptography, linguistics was also essential for developing effective language training courses for the army, and for translating foreign texts. During the Cold War that followed the World War, it was especially the feeling that quick translation of Russian research and intelligence into English was badly needed that prompted military investments into automatic translation. Warren Weaver's 1949 report on mechanical translation started off a decade of intensive work on automatic translation.

Though digital computing had its origins mainly in solving scientific and business computing, its redefinition in the late 1940s as 'information processing' had broadened the ambitions and the agenda of digital computing. Linguistic computing now became definitely a part of this enlarged vision. Even more, it has been contended that, as automatic translation was on the agenda, it contributed to spreading the use of the words 'translation' and 'language' among the communities of people that programmed and coded the early digital computers.[2] In any case, there was some convergence between the formal methods developed in automatic translation and those developed for automatic programming; in addition, one of the first string-processing languages, COMIT, was developed by Victor Yngve's team at M.I.T.

Research groups in automatic translation did not necessarily rely on already existing linguistic models and theories. Although the so-called neo-Bloomfieldians in U.S. linguistics had already developed a particular taste for formal and/or mathematical approaches to language (cf. Chap. 4), laying the foundations for what is now called 'immediate constituent analysis', their concepts did not transfer immediately to this new context. In many cases, ad hoc procedures were developed, often more determined by practical constraints and pragmatism than by insight into

[2] David Nofre, Mark Priestley, Gerard Alberts, 2014, 'When technology became language: The origins of the linguistic conception of computer programming, 1950–1960', Technology and Culture, Vol. 55, Nr. 1, p. 40–75.

language itself. For instance, ideas were borrowed from cybernetics and information theory and this transfer of concepts led to many discussions within linguistics itself, thereby changing the reference framework in which linguistics operated (cf. Chap. 5). One of the recurrent themes of debate was whether a mathematical approach to language was able to capture essential features of language, such as grammar or perhaps even meaning? And also, what kind of mathematics was best suited for formalising language: an algebraic, an analytical or probabilistic approach, or perhaps a combination of those?

One of the most important developments in modern linguistics, Noam Chomsky's generative grammars, grew out of this context. Chomsky came from neo-Bloomfieldian linguistics and was influenced in particular by Zellig S. Harris's work, he also worked for a while at the Mechanical Translation unit of M.I.T.'s Research Laboratory of Electronics. Though Chomsky systematically underplays the importance of this context, the encounter with computer people and mathematicians at M.I.T. and the discussions with other linguists around automatic translation and the mechanisation of language set the scene for his work. Furthermore, the mathematical precision that the French mathematician M.P. Schützenberger brought to Chomsky's models for language and the adoption of these models for computer-theoretical work on automata and programming were quite crucial in securing and augmenting the intellectual credibility and legitimation of his linguistic models. From 1962 onwards, Chomsky's theory would start its triumphant march in modern linguistics marking the success of the formalisation of language. Around the same time, though completely independently, disappointment about the poor results of automatic translation set in during the early 1960s. The very idea of automatic translation even got discredited as witnessed by the negative reports by Bar-Hillel (1960) and Automatic Language Processing Advisory Committee (ALPAC, 1964) (cf. Chap. 2).

Linguistics was also at the forefront of research in computerised information retrieval. Though the focus of much work in automatic translation was on the formal (syntactic) structure of language rather than the lexical structure, automatic translation's need for indexing and retrieving entries in a large, structured database such a dictionary or thesaurus stimulated research on automatic documentation and information retrieval (cf. Chap. 7). In the *Proceedings of the International Conference on Scientific Information. Washington DC November 16–21 1958* that brought various people from this emergent field together for the first time, researchers in automatic translation such as Victor Yngve, Anthony Oettinger, Margaret Masterman, or Zellig S. Harris figure alongside patent offices and (bio)chemists, who also had keen interest in automating large databases for storing, indexing and comparing their patents or (bio)chemical compounds. Automatic documentation together with automatic syntactic analysis would eventually be incorporated into the field of Natural Language Processing (NLP) in the 1990s (cf. Chap. 6).

Corpus linguistics, which had a long tradition in the empiricist English linguistics (Firth, Halliday), also invested itself slowly in the computerisation of linguistics. Though the first experiments were hampered by the limited computing and memory resources available at the time, the main concepts were refined over the

years and gradually bigger corpora were constituted and made amenable to automatic syntactic and lexical analysis. Matching the limited possibilities of the 1980s, some linguists shifted their focus to the study of 'specialised' languages, also called restricted languages (Firth, Halliday) or sublanguages (Harris). By the 1990s, powered by cheaper and faster computing and memory systems (namely, the revolution of the microcomputer), the integration of these tools and ideas led to today's Natural Language Processing (cf. Chap. 10).

***

This book is the outcome of a lifelong research interest in the history of the automation of language in the age of the digital computer. Jacqueline Léon, probably the world's foremost expert on this topic, moved from being a practitioner of the field to writing the history of the field. She worked as a CNRS researcher at the research institute for the history of linguistics, the Centre pour l'Histoire des Théories Linguistiques (UMR 7597, Paris). For her habilitation in 2010, she selected and reworked a number of her articles and also added some chapters to arrive at a quite complete, detailed history of the early automation of language, covering the latter half of the twentieth century. Her habilitation resulted into a book that was published as *Histoire de l'automatisation des sciences du langage* in 2015.[3] In her work much attention is paid to the context in which this encounter between linguistics and computing took place, and to the ensuing institutionalisation. While much of the important, and better-known, work was done in the United States, Léon's book also brings into focus equally important work that was done in Europe, notably in the United Kingdom. (Chaps. 7 and 10), France (Chaps. 8 and 9) and the Soviet Union (Chap. 7). Apart from some minor changes and additions, the present English translation is essentially faithful to the French original. However, the order of chapters has been changed to reinforce the chronological and logical structure of the book and a general conclusion has been added.

Université Paris 8
Saint-Denis, France

Maarten Bullynck,

[3] Jacqueline Léon, *Histoire de l'automatisation des sciences du langage*, ENS Éditions: Lyon, 2015. The book was reviewed in a number of journals:

- Ycard Bernard, 2015, Historiographia Linguistica 42 2–3: 457-460.
- Galazzi Enrica, 2016, L'analisi linguistica e litteraria 2016-1, p.173
- François Jacques, 2017, Bulletin de la Société de Linguistique 112-2: 1–9
- Bertrand Emanuel, 2018, 'Mathématisations et automatisations des sciences du langage : des tournants conceptuels ou technologiques?', Revue d'Histoire des Sciences Humaines, 32, pp.288–292.

Acknowledgements

I would like to express my sincere thanks to Maarten Bullynck for his help at several steps of the translation of the text, for his suggestion of modifying the order of the chapters, of adding footnotes to make the text more readable to non-linguists, and for the final edition of the manuscript.

Many thanks also to John DeWitt and Nick Riemer for their precious help with the English version.

Contents

Chapter 1
Introduction

The automation of the language sciences started with the first experiments on machine translation undertaken just after the end of World War II in 1948–1949 in the USA and in Great Britain. I will call it "the computational turn".[1] This turn, which defines how linguists adopted and integrated the concepts and the methods of the computer sciences and mathematics, involves very specific features. It is characterised by the sudden appearance of a set of completely new concepts, methods and practices which did not belong to the "horizon of retrospection" of linguists and the language sciences.

As it is affected by temporality, scientific activity refers to its past and its future. It refers to a body of knowledge which has been developed previously (concepts, works, results) and sets itself various projects for the future. Auroux (1987, 2007) names these two types of references "horizon of retrospection" and "horizon of projection", respectively. One of the tasks of the historian of sciences is to identify the structure of the horizon(s) of retrospection.

The horizon of retrospection is transmitted through tradition. As a body of previous knowledge, it can be common to a group and an area or specific to a researcher. The horizon of retrospection can be subject to contradictory moves. For a given domain, it can be subject to oblivion: some pieces of knowledge are regularly removed from the horizon of retrospection depending on the research interests of the moment. Alternatively, it can be subject to cumulation so that new knowledge can be aggregated with previous results. When emerging, a discipline reinvents its past renegotiating its horizon of retrospection by using these contradictory moves.

The horizon of retrospection initiated by machine translation is characterised by the intertwining of engineering with fundamental sciences such as mathematics, logic, physics, the neurosciences, acoustics and recently developed sciences like cybernetics and information theory. Developed primarily at MIT, this intertwining,

[1] See Rorty's *The Linguistic Turn* (1967) on the attention paid to the importance of language in the formulation of philosophical questions.

J. Léon, *Automating Linguistics*, History of Computing,
https://doi.org/10.1007/978-3-030-70642-5_1

named the war sciences by Dahan and Pestre (2004), led to cutting-edge technologies like radar, anti-aircraft defense systems and computers and later, after World War II, to machine translation.

In this context, the computerisation of the language sciences constitutes an interesting case because two horizons of retrospection confront one another: that of machine translation, namely, the horizon of retrospection of the war sciences in which linguistics has no place, and that of the linguists which, most of the time, does not involve the war sciences. Actually, the computational turn is marked at the beginning by a paradox, namely, that, although machine translation implies the (automatic) treatment of languages, linguistics does not belong to the war sciences.[2] Thus, for the language sciences, the new horizon was entirely new and set up by external way. However, because machine translation, and afterwards computational linguistics, deals with natural languages, the new field is essential to the language sciences, so that they have to integrate it.

The issues raised by the computerisation of the language sciences can be addressed by the following questions: how can the horizon of retrospection of machine translation be integrated by the linguists in their own horizon(s) of retrospection? Will this new knowledge cumulate with linguists' knowledge or will it replace the former horizon of retrospection, causing its cancellation or even its oblivion? Can it be renegotiated? What is claimed in this book, is that, in order to integrate a new horizon of retrospection (the body of knowledge common to machine translation pioneers), the linguists will tap into their own scientific and intellectual tradition in order to integrate this new set of knowledge. Knowing that, facing that issue, each tradition will operate differently.

Linguistics' computational turn happened in two steps. The first step, machine translation, while instituting a new horizon of retrospection, projected a future, a horizon of projection, for the language sciences. The second step, computational linguistics and the Chomskyan program, becomes the horizon of projection anticipating the future of the computational mathematisation of language. That period of 15 years (1948–1966), from the beginning of early machine translation experiments to the establishment of computational linguistics, can be regarded as constituting the computational turn.

The computational turn is anchored in "the first mathematisation of language" of the 1930s which can be defined by the rise of formalisation promoted by the Vienna School, in particular Carnap, as a common objective for every science, setting up mathematics as one language among others. The first mathematisation of language is characterised by the setting in interaction of algorithms and formal languages resulting from mathematical logic. The domain that rose at the junction of syntactic analysis, formal languages and computer programming can be called "the second mathematisation". Algorithms which were only an abstraction for logicians in the

[2] Martin Joos, an acoustic engineer and a phonetist, was an exception. He was probably the only linguist to have an activity in the war sciences (see Sect. 4.2.1).

first mathematisation came to be implemented in space and time. This is why it can be called the computational mathematisation of language.

The second mathematisation of language started with the implementation of syntactic analysis methods for machine translation which led to computational linguistics as an autonomous and institutionalised field of research. It can be claimed that it is thanks to machine translation, i.e. thanks to the strategic need for producing mass translations, that the formal languages, anchored in mathematical logic in the 1930–1940s, became involved in algorithms for syntactic analysis. This move determined the rise of formal grammars, in particular Chomsky's generative grammar.

A second key moment of the automation of the language sciences can be identified in the 1990s, when the rising power of computers and software would make much computerised textual data available. The appearance of microcomputers would lead linguists to use computerised data and new linguistic tools. This second turn which I call the "corpus turn" has characteristics quite different from the first turn. Actually, the use of large computerised corpora did not constitute a real departure from linguistic traditions. It allowed linguists to implement some assumptions belonging to earlier linguistic approaches, anterior to the first turn, making a link with earlier methods, like probabilistic methods resulting from information theory.

In this book, I will focus less on the social consequences of the automation of language (Auroux 1996) than on the various modes of integration (of the new horizon of retrospection) by the language sciences. I will tackle this issue through the following questions:

(i) Has only one form of mathematisation, i.e. logico-mathematics, prevailed in the automation of the language sciences, as the development of computational linguistics would suggest, or have other forms of mathematisation come into play?
(ii) The modes of integration of the new horizon of retrospection can be considered only in a comparative way. Several traditions are examined, American, British and French, and to a lesser extent, the Russian tradition from which the sources are less accessible. The choice of these traditions is not fortuitous but results from regarding machine translation as a war technology. The countries considered here are the "winners" of World War II; they were engaged in the Cold War where machine translation occupied a strategic place. Much more than others which would follow, and in a much more massive way, these states invested considerable means in machine translation. The question that remains to be answered is whether linguistic and intellectual traditions, still quite distinct at that period, determined different modes of integration, and how.[3]
(iii) I will examine how the new area opened by computational linguistics led to the emergence of natural language processing and artificial intelligence.
(iv) I will investigate to what extent the possibility of automation gave rise to new objects and new methods in the language sciences. We will see that, thanks to

[3] See Léon (2014).

automation, lexical semantics will be renewed starting from old questions about the "word" as a linguistic unit.

(v) We will ask whether all the concepts and methods coming from the war sciences were integrated into the language sciences or whether some of them were privileged leaving aside others, and how. One thinks in particular of information theory, a central, unifying and universalising theory, which, at the time of its integration, knew various destinies, distinct from that of computational linguistics.

(vi) Another question concerns periodisation. We will ask whether, starting from the event constituted by the computational turn, a linear periodisation of integration can be delimited with a beginning and an end; or whether, on the contrary, various modes of integration will determine various periodisations, some of them anchored in previous centuries and still in progress today.

(vii) Lastly, we can ask whether this technological revolution constitutes a third revolution of the language sciences comparable to the emergence of writing and the grammatisation of vernaculars, respectively, the first and second revolution defined by Auroux (1994).

This book aims to give an account of three movements, machine translation as the founding event of the computational turn, the integration of this turn by the language sciences and the corpus turn. These three movements will be developed in nine chapters.

Five chapters are devoted to the USA, where everything started.

Chapter 2 "Machine Translation as Technology of War" gives an account of the event constitutive of the computational turn.

Chapter 3 "The War Effort, the Technologisation of Linguistics and the Emergence of Applied Linguistics" deals with the technologisation of the language sciences. It is devoted to the war effort undertaken by the Americans in the domain of language teaching, in which most American linguists were involved. Many were also implied in cryptography, the majority as simple translators of the messages written in "exotic" languages, but some taking part in the decoding work itself. The war effort led to the emergence of applied linguistics in the USA which was characterised by an important technologisation of the methods.

Chapter 4 deals with the automation of American linguistics. Entitled "The Computational Turn and Formalisation in Neo-Bloomfieldian Distributionalism", it examines how computationability raised new challenges for American structuralist linguists concerning translation and formalisation.

Chapter 5 "Information Theory: The Transfer of Terms, Concepts and Methods" is concerned less with the automation than with the mathematisation of language. I will examine the process by which some concepts and methods of information theory, bringing together telecommunication engineering and mathematical theories, were integrated into the language sciences. Roman Jakobson's Distinctive Feature Theory presents an exemplary case of this process associating European linguistics and engineering and American war sciences.