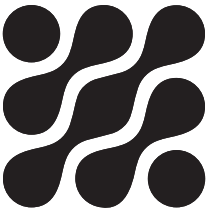
Lars Engwall, Tina Hedmo & Olle Persson

Corpus linguistics in Sweden

PIONEERS AND THEIR CONTEXT

KUNGL. VITTERHETS HISTORIE OCH ANTIKVITETS AKADEMIEN
HANDLINGAR
FILOLOGISK-FILOSOFISKA SERIEN 25

Lars Engwall, Tina Hedmo & Olle Persson



Corpus linguistics in Sweden

PIONEERS AND THEIR CONTEXT

KUNGL. VITTERHETS HISTORIE OCH ANTIKVITETS AKADEMIEN
HANDLINGAR
FILOLOGISK-FILOSOFISKA SERIEN 25

Engwall, L., Hedmo, T. & Persson, O., *Corpus linguistics in Sweden: Pioneers and their context*. Kungl. Vitterhets Historie och Antikvitets Akademien (KVHAA), *Handlingar, Filologisk-filosofiska serien* 25. Stockholm 2019. 138 pp.

Abstract
This volume presents findings from research on the development of corpus linguistics in Sweden as a scientific innovation. It begins with a presentation of the early international development of corpus linguistics as well as the institutional and disciplinary conditions for research on the subject in Sweden, followed by accounts of the first generations of Swedish innovators. External funding and international development were important for these pioneers, alongside the fact that established professors in language departments seem to have been relatively open to the new ideas. The criticism levelled against corpus linguistics appears instead to have come mainly from departments of general linguistics. In the course of time, this negative attitude has diminished and corpora have become an almost indispensable tool in linguistic research. Developments in Sweden are placed in an international perspective by means of an analysis of the publication database SciVerse Scopus for 1970–1999. It shows a research field in two well-defined clusters: corpus builders and corpus users, the former often with a background in language studies, the latter evincing considerable representation of psychologists and scholars of cognitive science with an interest in language acquisition and language loss. Evidence of the significance of international developments for scientific innovation is provided by an analysis of the development of professional organizations on both sides of the Atlantic.

Keywords
corpora, linguistics, scientific innovations, disciplinary conditions, institutional conditions, research funding, innovators, computers, lexicography, dictionaries, language acquisition, frequency studies

ISBN 978-91-88763-02-0
ISSN 0083-677X

Publisher Kungl. Vitterhets Historie och Antikvitets Akademien (KVHAA, The Royal Swedish Academy of Letters, History and Antiquities)
Box 5622, SE-114 86 Stockholm
http://www.vitterhetsakademien.se
Distribution eddy.se ab, Box 1310, SE-621 24 Visby
http://vitterhetsakademien.bokorder.se
Graphic design and cover Niklas Lindblad, Mystical Garden Design
Printed by Bulls Graphics, Halmstad, Sweden 2019

CONTENTS

CHAPTER 1. INTRODUCTION

Background

Human communication in written as well as spoken form has long interested scholars all over the world. One classical approach has been the collection of examples of expressions in order to analyse variations in constructions, dialects, etc. This empirically grounded approach contrasts with deductive approaches, i.e. the construction of theoretical examples and testing them in practice. Interestingly enough, both approaches experienced significant changes in the early decades after the Second World War. The development of computers then dramatically provided new opportunities to handle large bodies of text in a more systematic way. At the same time the introduction of the theory of generative grammar by Noam Chomsky (1957 and 1965) had a significant impact on linguistic research. As a result, the 1960s brought considerable tensions between empirically and theoretically oriented linguists. This happened all over the world, but more so in countries which were strongly influenced by developments in the United States. Sweden belongs to this group, and it did indeed exhibit these tensions. Nevertheless, as will be evident in this volume, Swedish scholars turned to corpus linguistics in the 1960s. Their choice of approach was not always accepted and was particularly questioned by those who had joined the Chomskyan camp. More than fifty years later we can note that corpus linguistics has become strongly established in linguistic research and is providing new opportunities in other areas as well. This has been demonstrated within a European comparative project, where corpus linguistics was chosen as one of four scientific innovations that were studied.

The background to the study was an invitation in 2008 from the European Science Foundation for proposals within a research programme on higher education.[1] Among the projects that were approved was 'Re-Struc-

1 The title of the programme was 'Higher Education and Social Change' (EuroHESC) and it

turing Higher Education and Scientific Innovation' (RHESI), for which Professor Richard Whitley at Manchester Business School was the main proponent. The application contained five research teams based in Germany, the Netherlands, Sweden, Switzerland and the United Kingdom. For Sweden, a group at the Uppsala University Department of Business Studies took part in preparing the application and in undertaking the research with the support from the Swedish Research Council.[2] Research grants were likewise obtained from national funding bodies in Germany, the Netherlands and Switzerland, but unfortunately not in the United Kingdom. The project could therefore only cover four countries, for which the research team decided to study four scientific innovations, two in the Natural Sciences: (1) Bose-Einstein Condensation (BEC), and (2) Evolutionary Developmental Biology (Evo-Devo) and two in the Humanities and the Social Sciences: (3) International Large Scale Student Assessments (ILSA), and (4) Corpus Linguistics (CL). The research design and the output can therefore be summarized as in Table 1.1, which also shows the focus of the present volume, that is, corpus linguistics in Sweden.

The research has been based on available literature as well as interviews with selected individuals in the four fields in the four countries.[3] Results have been presented in an edited volume (Whitley & Gläser, 2015), which has provided comparative analyses across countries for the four innovations: see Laudel et al. (2015a) for Bose-Einstein Condensation (BEC), Laudel et al. (2015b) for Evolutionary Developmental Biology (Evo-Devo), Gläser et al. (2015) for International Large Scale Student Assessments (ILSA), and Engwall et al. (2015) for Corpus Linguistics (CL). The latter paper has constituted an important input for the present publication. This has also been the case with a paper where the organizational development of scientific fields is analysed with evidence from the field of corpus linguistics (Engwall & Hedmo, 2016).

---

was part of the EUROpean COllaborative RESearch (EUROCORES) scheme.

2 Grant 90671701, which is acknowledged with gratitude.

3 For the interviewees in the Swedish project, see p. 15.

Table 1.1. Research design and output

Scientific innovation	Output	Country studies			
		Germany	Netherlands	Sweden	Switzerland
Bose-Einstein Condensation (BEC)	Laudel et al. (2015a)				
Evolutionary–Developmental Biology (Evo-Devo)	Laudel et al. (2015b)				
International Large Scale Assessments (ILSA)	Gläser et al. (2015)				
Corpus Linguistics (CL)	Engwall et al. (2015)			The present volume	

A model for analysis

The above-mentioned joint article on corpus linguistics in the four countries (Engwall et al., 2015) demonstrates how corpus linguistics started in the 1960s in three of the four countries studied: Germany, the Netherlands and Sweden, while it was not developed in Switzerland until the recruitment of foreign linguists in the 1990s. And, although the Netherlands had corpus linguistics as early as in the 1960s, progress was slower there than in Germany and Sweden. For Germany there is no doubt that the creation of the Institute for the German Language (*Das Institut für Deutsche Sprache*,

IDS) was very important for the advance. In Sweden, on the other hand, it was instead a combination of an academic entrepreneur, international influences and funding from a variety of agencies that lay behind the early projects. Interestingly enough, as we will show in Chapters 5 and 6, an institution similar to IDS developed in Gothenburg. However, this was rather a bottom-up than a top-down project.

The slower adoption of computer linguistics in the Netherlands and Switzerland seems to have been the effect of stronger alternative research communities, namely generativists, and, in Switzerland, strong groups in historical linguistics. It is also probable that the later adoption of corpus linguistics in Switzerland could be due to the fact that the country has four official languages, in contrast to the others, which have one dominant language each. The pioneers in these countries thus started out with the majority language, while in Switzerland it was English that was chosen for the early corpora, not one of the country's official languages.

Generally speaking, a major force behind the development of corpus linguistics was the advance of computer technology. At the same time, however, it should be pointed out that important individual pioneers in the United Kingdom and United States provided a powerful impetus. These individuals, in turn, inspired academic entrepreneurs, most of them men in their early careers.

On the basis of the above observations we were able to formulate a model regarding the conditions that influence the behaviour of scientific entrepreneurs, that is, the individual actors who pursue new avenues in their research. Two types of conditions were found to be significant: *institutional conditions* and *disciplinary conditions* (Figure 1.1).

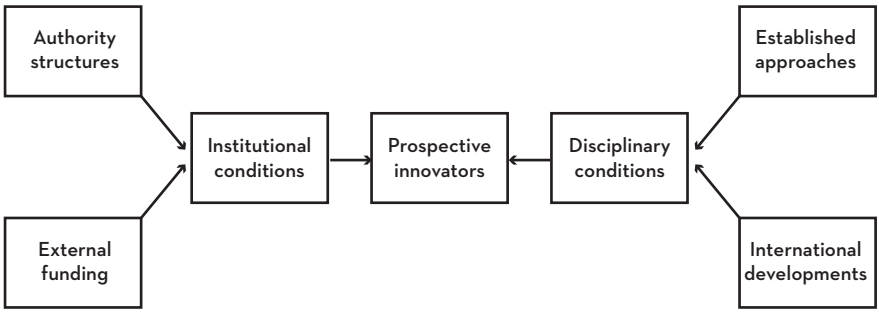


Figure 1.1. Conditions for prospective innovators.

The institutional conditions (left-hand side of Figure 1.1) are highly dependent on *authority structures*, meaning the extent to which established professors have the prerogative and willingness to control the scientific activities of their younger colleagues. If this control is strong, we may expect innovations to be hampered, while the opposite is true in cases where an open atmosphere prevails. Needless to say, the opportunities of control are stronger the more the established professors control critical resources. Therefore, we can expect the availability of *external funding* to diminish the effects of this control.

The disciplinary conditions (right-hand side of Figure 1.1) are constituted by the specific settings of a scientific field. Central are *the established approaches* (or paradigms in the vocabulary of Thomas S. Kuhn, 1962), which vary with the degree of task uncertainty and the dependence between researchers in the field (Whitley, 1984). However, they may also vary across different geographical areas, despite the fact that research has long been an international activity. At the same time, the latter circumstance implies that even if national gurus try to restrict their country's research to their own favourite approaches, *international developments* are likely to counterbalance conservative forces and successively influence the institutional conditions in other directions.

Outline of the volume

On the basis of the described model, Chapter 2 starts out by summarizing early international developments in corpus linguistics. Chapters 3 and 4 recapitulate the Swedish institutional and disciplinary conditions, respectively. In Chapter 5 a first generation of Swedish corpus linguists is presented, while Chapter 6 deals with two scholars from the second generation of Swedish corpus linguists working with written language. Similarly, Chapter 7 presents later corpus linguists focusing on spoken language, while Chapter 8 provides evidence regarding later international developments by means of a bibliometric analysis of publications during the period 1970–2010 as well as the organizing of the field. The overall conclusions are given in Chapter 9.

Appendix: List of interviewees for this volume

Interviewee	Date	Department	University	Born	Interviewer
Gunnel Engwall	110419	French	Stockholm	1942	Tina Hedmo
Bernard Quemada	110509	French	Besançon	1926	Gunnel Engwall
Robert Martin	110511	French	Paris	1936	Gunnel Engwall
Jussi Karlgren[4]	110623	Speech technology	KTH	1965	Tina Hedmo
Inger Rosengren	110818	German	Lund	1934	Lars Engwall
Åke Viberg	110829	Linguistics	Uppsala	1945	Tina Hedmo
Lars Borin	110908	Swedish	Gothenburg	1957	Tina Hedmo
Jens Allwood	111001	Linguistics	Gothenburg	1947	Tina Hedmo
Merja Kytö	111109	English	Uppsala	1953	Tina Hedmo
Sture Allén	111117	Swedish	Gothenburg	1928	Lars Engwall
Jan Svartvik	111202	English	Lund	1931	Lars Engwall
Geoffrey Leech	130509	English	Lancaster	1936	Lars Engwall

4 Karlgren is adjunct professor at the Royal Institute of Technology (KTH). His main employer is the text analyst company Gavagai.

CHAPTER 2. EARLY INTERNATIONAL DEVELOPMENTS

Introduction

Corpus linguistics, the focus of this volume, concerns studies of language within defined bodies (collections) of text. This approach to language studies has long traditions, long before the term corpus linguistics was coined in the early 1980s.[5] It is based on the idea that studies of language have to be based on a systematic compiling of written and spoken language. Before the advent of computers this was mainly accomplished through the visual scanning of selected texts for the identification of word use and expressions. Obviously, the development of information technology has changed the conditions for such studies considerably. However, it is very important to keep in mind that the conditions for the early users of computers were significantly different from those in the early twenty-first century. The early computers were slow, had rather restricted memory capacity and were more suited to handling mathematical calculations than texts. Over time conditions have changed dramatically through the development of both hardware, that is, much faster computers with extensive memory capacity, and software in terms of computer programs for the treatment and analysis of written as well as spoken language. In this way, modern linguists have access to a vast number of comprehensive language databases. This in turn has paved the way for what is more and more being called digital humanities. The use of these large-scale databases is not limited to scholars in the humanities, however. They are also used by researchers in

5 According to McCarthy & O'Keefe (2010, p. 5) Aarts & Meijs (1984) 'is seen as the defining publication as regards coinage of the term'.

other fields, such as medicine and psychology. Obviously, corpus linguistics has also had significant implications for the development of information technology itself in applications such as spelling programs, voice recognition and search algorithms. In this way there is a dialectical relationship between corpus linguistics and information technology. The advent of this development occurred in the wake of the Second World War through the work of a group of scientific entrepreneurs who took advantage of emerging computer facilities. In this way they paved the way for modern language studies as well as language-related research and applications in other fields.

This chapter will first present the international pioneers. It will then discuss forces working in favour of and against corpora, provide the results of a bibliometric analysis of the international roots of corpus linguistics, and finally present conclusions.

International pioneers

Internationally, scholars of languages have long used corpora for the production of dictionaries, dialect atlases and grammars. A very early example is a German frequency dictionary (Kaeding, 1897–1898), produced by Friedrich Wilhelm Kaeding (1843–1928), an expert in stenography. Other early examples are Henmon (1924) and the publications of the American and Canadian Committees on Modern Languages (cf. e.g. Vander Beke, 1929; Buchanan, 1931; Cheydleur, 1934; Morgan, 1933). In the 1930s, studies such as these inspired the Harvard linguist Professor George Kingsley Zipf (1902–1950) to formulate what has become known as Zipf's law, which states that the product of rank and frequency in word distributions tends to be constant (Zipf, 1932).

Later on, in the 1950s, the Italian Jesuit Pater Roberto Busa (1913–2011) made early contributions through his work to provide concordances of

the texts of Thomas Aquinas (cf. e.g. Busa, 1951).[6] One of Busa's students, Antonio Zampolli (1937–2003), subsequently became a very active scholar in the field of computational linguistics (cf. e.g. Atkins & Zampolli, 1994), not least through the Pisa Summer Schools in the 1970s and the creation of the Pisa Institute of Computational Linguistics.[7]

Among European pioneers the Frenchman Bernard Quemada (1926–2018) can be taken as an illustrative case for the conditions the pioneers faced.[8] He started his work on computational linguistics in the 1950s in Besançon. Thanks to a considerable faculty grant and contacts with the French computer company Bull, and despite resistance from older colleagues, he was able to create a laboratory for the study of French vocabulary.[9] In this work the occurrence of accents in French created particular problems, which were eventually solved in collaboration with the computer company IBM. Quemada approached the then rector of the Academy of Nancy, the linguist and lexicographer Paul Imbs (1908–1987), who in 1960 had founded the French National Institute of the French Language (*l'Institut National de la Langue Française*, INaLF) in Nancy for the development of French lexica.[10] Quemada managed to convince Imbs of the advantages of using electronic data processing. During the period from 1959 to 1993 he edited thirty volumes presenting historical French vocabulary (Quemada, 1959–1993) and defended his thesis on dictionaries of modern French in 1968 (Quemada, 1968). He worked as deputy director of INaLF and became its

---

6 For an obituary, see http://www.guardian.co.uk/higher-education-network/blog/2011/aug/12/father-roberto-busa-academic-impact (accessed on July 28, 2017).

7 See http://www.mt-archive.info/LREC-2004-Zampolli.pdf (accessed on July 28, 2017) and Johansson (2008, p. 35).

8 This paragraph is based on a personal interview with Bernard Quemada by Gunnel Engwall on May 9, 2011.

9 Incidentally, Bernard Quemada got the idea to use punched cards for his language studies by observing a service man from the electricity company using such cards for registering meter readings.

10 In 1957 Paul Imbs had arranged a colloquium that paved the way for later developments (see CNRS, 1961).

director in 1977.[11] He remained in this position until 1992, when he moved to Paris, succeeded as director by Robert Martin (b. 1936). At an early stage Quemada arranged summer schools, which attracted students like Antonio Zampolli, and the Manchester scholar Peter Wexler (1923–2002).[12] Among faculty members were the grand old man of French frequency studies Charles Muller (1909–2015).[13] Bernard Quemada's significance for the field is evidenced by a Festschrift in two volumes (Zampolli, Cignoni & Peters, 1981). Apparently independently of Europe-based researchers, the Rumanian-born Stanford professor Alphonse Juilland (1923–2000) produced frequency dictionaries of the four Romance languages Spanish (Juilland & Chang-Rodriguez, 1964), Rumanian (Juilland, Edwards & Juilland, 1965), French (Juilland, Brodin & Davidovitch, 1970) and Italian (Juilland, Traversa & Beltramo, 1973).

Although Busa, Quemada and Juilland appear to have been forerunners, the literature often points to Henry Kučera (1925–2010) and Nelson Francis (1911–2002), the creators of the Brown corpus at Brown University in Providence, RI, as the pioneers. Their corpus contained around one million words that had been published in the United States in 1961. It was analysed and published as *Computational Analysis of Present-Day American English* in 1967 (Kučera & Francis, 1967). The corpus later on provided the basis for the publication of the first edition of *The American Heritage Dictionary* in 1969.[14] The Brown corpus was no doubt an inspiration for many followers in the field of corpus linguistics. The closest follower was the CAMET project (Computer Archive of Modern English Texts), launched in 1970 by the then reader in English at Lancaster University, Geoffrey Leech

---

11 The work at INaLF provided the basis for *Le Trésor de la Langue Française Informatisé* (TLFi), which is a dictionary of the French language available on-line, CD and as books (*Trésor de la langue française informatisé*, 2004).

12 For Wexler's Festschrift, see Durand (1983).

13 Cf. e.g. Muller (1967, 1968 and 1979). For the Festschrift at the celebration of Muller's centenary, see Delcourt & Hug (2009).

14 For the outcome of their later work, see Francis & Kučera (1982).

(1936–2014).[15] Targeting British English, it was collected according to the same principles as the Brown corpus.[16] In time, through collaboration with Norwegian scholars, particularly Jan Svartvik's student Stig Johansson, it became the Lancaster-Oslo/Bergen (LOB) corpus and was completed in 1978 (Johansson, 2008).[17] Another initiative worth mentioning is that of the London professor Randolph Quirk (1920–2017), who launched the project Survey of English Usage (SEU) at University College London as early as 1959.[18] In so doing, he turned to the collection not only of written texts but also of spoken English (cf. Quirk & Svartvik, 1978 and further below in Chapter 5, pp. 52–54).

In Germany Hans Eggers (1907–1988) took an early initiative in 1956 at the University of Saarland. However, it was not until 1968 that the corpus consisting of 200,000 words of German text was completed. In the meantime, in 1964, the above-mentioned Institute for the German Language (IDS) had been founded in Mannheim by the federal and provincial governments to study and document the 'contemporary usage and recent history of German language'. The first outcome of this initiative was a newspaper corpus (*Das Bonner Zeitungskorpus*) of 3.1 million words compiled by Manfred W. Hellmann (b. 1936).[19] A second one was the Freiburger Korpus of

---

15 For his Festschrift, see Thomas & Short (1996).

16 According to Geoffrey Leech, he got a very positive answer from Nelson Francis, when asking the question 'What do you think about the idea of a British corpus to match the Brown corpus?': 'Yes, and for heaven's sake, make it as close a match as possible so that comparisons can be made.' (Interview with Lars Engwall May 9, 2013.)

17 The year before the LOB corpus was completed (1977) the International Computer Archive of Modern English (ICAME) had been founded by five key researchers, among them Nelson Francis, Geoffrey Leech, Stig Johansson and Jan Svartvik. The purpose of this organization was to assemble all available English corpora (http://icame.uib.no/history/founding_document_1977.pdf, accessed on July 28, 2017, see further Chapter 8, p. 89). A significant reason for the founding of ICAME was the need to put pressure on publishers to give permission to use the selected texts in the LOB corpus. (Geoffrey Leech in interview with Lars Engwall, May 9, 2013.)

18 According to Geoffrey Leech, Randolph Quirk's work was supported by the publisher Longmans. (Interview with Lars Engwall, May 9, 2013.)

19 See Eggers (1969).

spoken standard German, started in 1968 by Hugo Steger (1929–2011).[20] These corpora were followed by several others within IDS.[21]

Forces working for and against corpora

It is apparent that the development of computer technology was important for the development of corpus linguistics. However, there are also reasons to point to the fact that the 1960s also brought a questioning of the collection of vast databases. Hence, Fillmore (1992, p. 35) has described this as the tension between 'armchair linguists' and 'corpus linguists'. And, although corpora spread, according to Johansson (2008, p. 33) 'the negative view of corpora found in early generative linguistics persisted in many circles'.

As mentioned, the MIT linguist Noam Chomsky (b. 1928) was the key person in this context with the idea of the transformational grammar (Chomsky, 1957 and 1965). The important distinction in his theory was that between *competence* (the language knowledge of a native speaker) and *performance* (the language used).[22] As a consequence he and his followers argued that it would be more appropriate to study language by confronting native speakers with constructions rather than by collecting vast materials of written and spoken language. In this way corpus linguistics was to a large extent challenged by general linguistics.[23] The Chomsky approach certainly

---

20 See *Gesprochene Sprache* (1974).

21 See further Engwall et al. (2015), pp. 339–342.

22 It should be noted that as early as the beginning of the last century the Swiss structural linguist Ferdinand de Saussure (1857–1913) made a similar distinction between *langue* (the grammar) and *parole* (the spoken language and the written text) (see further Saussure, Bally & Sechehaye, 1916). This structuralist approach was challenged by Chomsky, however.

23 In the words of Chomsky (1957, p. 159): 'Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite.' And, according to Geoffrey Leech, Robert Lees, a supporter of Chomsky, told Nelson Francis, when he heard about the plans to create the Brown corpus: 'Corpus? What a complete waste of time. In five minutes I could supply you with more examples from my head than you can find in the whole Library of Congress.' (Interview with Lars Engwall, May 9, 2013.)

had the advantage of requiring fewer resources and better opportunities for the publication of articles in international journals. However, it has also been subject to criticism.[24]

While the Chomsky approach challenged computational linguistics, commercial forces were working for the creation of large databases. As mentioned above the Brown Corpus became the basis for a new dictionary of American English. Likewise, other publishers took a similar interest, including Oxford University Press (OUP), which collaborated with the Arts Computing Centre at Waterloo, Ontario, for the creation of the *Oxford Dictionary of English* (Johansson, 2008, p. 35). This led to the creation of the British National Corpus, which is an industrial/academic consortium led by OUP funded by commercial partners as well as the British government, now containing 100 million words.[25] Needless to say, the development of this as well as other corpora has strongly been facilitated by changes in printing technology since the 1970s leading to easy access to the content in newspaper articles, books and other publications.

Another force in favour of corpus linguistics was the efforts to use computer technology for translation. Thus, as early as 1962 the Association for Machine Translation and Computational Linguistics (AMTCL) was founded for 'the international scientific and professional society for people working on problems involving natural language and computation', which in 1968 took its present name the Association for Computational Linguistics (ACL).[26] At the same time research centres for computer analysis were created on both sides of the Atlantic, for example at the University

---

24 For a Swedish example, see Öhman (2007).

25 See www.natcorp.ox.ac.uk/ (accessed on July 28, 2017). For the development of the *Harper Collins Dictionary*, see Sinclair (1987). With respect to the latter, John Sinclair and his group in Birmingham were, according to Geoffrey Leech, less interested in grammar and semantics than the Lancaster group and instead focusing on co-location of words. (Interview with Lars Engwall, May 9, 2013.)

26 See http://www.aclweb.org/archive/misc/History.html, accessed on July 28, 2017. On the organizing, see Chapter 8, p. 89.

of California, Irvine (*Thesaurus Linguae Graecae*), and the universities of in Bergen, Bonn, Mannheim, and Saarbrücken (Johansson, 2008, p. 35).

In relation to the tensions between the supporters of Chomsky and corpus linguists, it is also important to bear in mind that not all linguists deal with present-day language, which permits interaction with native speakers. A prime example of this is Father Busa and his studies of Thomas Aquinas mentioned above. The same is true for studies of medieval languages, for instance. Therefore, the former director of INaLF, Robert Martin, has thus denied in an interview any critical attitudes towards his corpus work.[27]

The international roots of corpus linguistics

In order to further map the international roots of corpus linguistics, the database SciVerse Scopus was searched within the project in August 2010 using the following search algorithm:[28]

> ALL ("corpus linguistics" OR "word frequencies" OR "frequency dictionary" OR "computational lexicology" OR "statistique lexicale" OR "vocabulaire" OR "frequenzwörterbuch" OR "statistique linguistique" OR "häufigkeitswörterbuch" OR "dictionnaire des frequencies" OR "ordfrekvenser" OR "frekvensordbok" AND (LIMIT-TO(SUBJAREA, "ARTS")).

The search resulted in 3,967 articles and reviews. When the cited references

---

27 Interview with Robert Martin, by Gunnel Engwall on May 11, 2011.

28 The search was performed by Professor Olle Persson, Inforsk, Umeå University, Sweden, and was made in all fields including cited references. SciVerse Scopus is the world's largest abstract and citation database of peer-reviewed literature and quality web sources. According to its website it is 'the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings'. In July 2017 it covered 67 million records from some 22,000 peer-reviewed journals (https://www.elsevier.com/solutions/scopus, accessed on July 28, 2017).

were divided into the four periods of 1900–1939, 1940–49, 1950–59 and 1960–69 (Table 2.1), a number of well-known works appeared.

As for the 1900–1939 period (Table 2.1, first section) we can first note the above-mentioned George Zipf and his *The Psycho-Biology of Language* (Zipf, 1935), and two structuralists, Leonard Bloomfield (1887–1949) with *Language* (1933) and Ferdinand de Saussure and collaborators with *Cours de linguistique générale* (Saussure, Bally & Sechehaye, 1916). However, there are also links to the classical languages through two books dealing with Greek (Schwyzer, 1939; Chantraine, 1933) and one (Ernout & Meillet, 1932) with Latin.

During the second period (Table 2.1, second section) Zipf is still a frontrunner, this time with his *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Zipf, 1949), followed by the ground-breaking paper on information theory by Claude Shannon (1916–2001), 'A Mathematical Theory of Communication' (Shannon, 1948) as well as a co-authored book by Edward Thorndike (1874–1949) and Irving Lorge (1905–1961) for educational purposes: *The Teacher's Word Book of 30,000 Words* (Thorndike & Lorge, 1944). They are followed by G. Udny Yule (1871–1951), a well-known statistician who published his *The Statistical Study of Literary Vocabulary* (Yule, 1944) during the Second World War. Last among the frequently cited works from the 1940s are one book on neuropsychology (Hebb, 1949) and another on the names in Indo-European languages (Benveniste, 1948). Clearly, the works in the second period point to the interdisciplinary character of the emerging field.

The top reference from the 1950s (Table 2.1, third section) is the English linguist John Rupert Firth (1890–1960), who after a decade at the University of Punjab returned to London, where he became Professor of General Linguistics. His *Papers in Linguistics 1934–1951* (Firth, 1957) is followed by an educationally oriented volume, *A General Service List of English Words* (West, 1953) and a dictionary, *Indogermanisches etymologisches Wörterbuch, Bd 1* (Pokorny, 1959) compiled by the Austrian linguist Julius Pokorny (1887–1970).

Among the following titles, Noam Chomsky's *Syntactic Structures*

Table 2.1. The most cited works from 1900–1939, 1940–1949, 1950–1959 and 1960–1969 in a SciVerse Scopus search for corpus-related works

1900–1939

Zipf, George Kingsley, 1935, The Psycho-Biology of Language: An Introduction to Dynamic Philology. Boston: Houghton Mifflin Company.
Bloomfield, Leonard, 1933, Language. New York: Holt, Rinehart and Winston.
Schwyzer, Eduard, 1939, Griechische Grammatik. Bd 1, Allgemeiner Teil, Lautlehre, Wortbildung, Flexion. München: Beck'sche Vlgs-Buchhandlung.
Chantraine, Pierre, 1933, La formation des noms en grec ancien. Paris: Champion.
Saussure, Ferdinand de, Charles Bally & Albert Sechehaye, 1916, Cours de linguistique générale. Lausanne: Payot.
Ernout, Alfred & Antoine Meillet, 1932, Dictionnaire étymologique de la langue latine. Paris: Klincksieck.

1940–1949

Zipf, George Kingsley, 1949, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, MA: Addison-Wesley.
Shannon, Claude, 1948, 'A Mathematical Theory of Communication', The Bell System Technical Journal, 27 (3 and 4), pp. 379–423 and 623–656.
Thorndike, Edward L. & Irvin Lorge, 1944, The Teacher's Word Book of 30.000 Words. New York: Teacher's College, Columbia University.
Yule, G. Udny, 1944, The Statistical Study of Literary Vocabulary. Cambridge: Cambridge University Press.
Hebb, Donald Olding, 1949, The Organization of Behavior: A Neuropsychological Theory. New York: Wiley.
Benveniste, Émile, 1948, Noms d'agent et noms d'action en indo-européen. Paris: Adrien-Maisonneuve.

1950–1959

Firth, John Rupert, 1957, Papers in Linguistics 1934–1951. London: Oxford University Press.
West, Michael, 1953, A General Service List of English Words, with Semantic Frequencies and a Supplementary Word-list for the Writing of Popular Science and Technology. London: Longman.
Pokorny, Julius, von, 1959, Indogermanisches etymologisches Wörterbuch, Bd 1. Bern: Francke.
Chomsky, Noam A., 1957, Syntactic Structures. New York: Mouton.
Berko, Jean, 1958, 'The Child's Learning of English Morphology', Word, 14 (2–3), pp. 150–177.
Miller, George A., 1956, 'The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information', Psychological Review, 63 (2), pp. 81–97.

1960–1969

Kučera, Henry & Nelson W. Francis, 1967, Computational Analysis of Present-Day American English. Providence, RI: Providence University Press.
Benveniste, Émile, 1969, Le vocabulaire des institutions indo-européennes, tome 1: Économie, parenté, société. Paris: Les éditions de Minuit.
Oldfield, Richard C. & Arthur Wingfield, 1965, 'Response Latencies in Naming Objects', Quarterly Journal of Experimental Psychology, 17 (4), pp. 273–281.
Chantraine Pierre, 1968, Dictionnaire étymologique de la langue grecque, tome 1, Paris: Klincksieck.
Morton, John, 1969, 'Interaction of Information in Word Recognition', Psychological Review, 76 (2), pp. 165–178.
Chomsky, Noam A. & Morris Halle, 1968, The Sound Pattern of English. New York: Harper & Row.

(Chomsky, 1957) is particularly worth noting, since it represents, as mentioned above, a different approach than corpus linguistics. The last two papers from the 1950s have a more psychological bent. The first (Berko, 1958), by the Boston University psycholinguist Jean Berko (b. 1931, Berko Gleason after marriage in 1959), focuses on language learning, while the second (Miller, 1956), by the then Harvard professor George A. Miller (1920–1992), deals with human information processing. This means that the top references in the 1950s came both from linguistics and psychology.

In the 1960s (Table 2.1, bottom section) the work of Henry Kučera and Nelson Francis (1967) is at the top, an indication of their significance as forerunners in corpus linguistics. However, there is also a structural linguist, Émile Benveniste (1902–1976) in second place with his *Indo-European Language and Society* (Benveniste, 1969). He is followed by two Oxford psycholinguists Richard Oldfield (1909–1972) and Arthur Wingfield (b. 1937), with their paper 'Response Latencies in Naming Objects' (Oldfield & Wingfield, 1965). In addition, we find another French title: the Greek dictionary *Dictionnaire étymologique de la langue grecque* (Chantraine, 1968), published by the Paris linguist Pierre Chantraine (1899–1974) as well as a paper by the British cognitive scientist John Morton (b. 1933) on word recognition (Morton, 1969). Last of the top works from the 1960s is the co-authored *The Sound Pattern of English* (Chomsky & Halle, 1968) by Noam Chomsky and Morris Halle (1923–2018). Again, we note the varied sources for the field of corpus linguistics: the results of corpus studies, studies of classical language, psychology and even the works of Noam Chomsky.

Conclusions

As shown above, corpus linguistics has its roots before the Second World War. As a matter of fact, such work was done as early as the late nineteenth century. However, the development of computer technology after the Second World War implied a major change in the conditions for language research. Thus, internationally a number of relatively young men – most

of them in their thirties, with Nelson Francis as an exception, having passed fifty – saw the opportunities with the new technology, managed to attract resources and were prepared to invest their time in building corpora. However, we have also seen from our SciVerse Scopus search that the efforts in corpus linguistics had roots in a mix of various earlier works from structuralism, statistics, information theory, education and psycholinguistics. Corpus linguists even had a large number of citations to the works of Chomsky.

CHAPTER 3. INSTITUTIONAL CONDITIONS IN SWEDEN

Authority structures

The Swedish system for research is closely related to the rules for universities and other institutions of higher education, since by tradition Sweden has a very small sector of research institutes. This is based on a strong belief in the Humboldtian principle of combining research and teaching. At the time of the early innovations in corpus linguistics in the 1960s almost all of the universities were public, the Stockholm School of Economics being the only private institution.[29] As for the authority structures two aspects are relevant for our analysis: (1) the structure of institutions, and (2) the power relations inside institutions.

The structure of institutions

The Swedish university system goes back to the late fifteenth century when Uppsala University was created by papal bull in 1477. It was followed by a second university in southern Sweden through the foundation of Lund University in 1666. These two universities were the only ones until the late nineteenth century, when two local university colleges were created, one in Stockholm in 1878 (upgraded to a state university in 1960) and the other in Gothenburg in 1891 (upgraded to a state university in 1954). A few decades after the Second World War universities were also created in Umeå in 1965 and in Linköping in 1975.[30] As will be evident below, the above-mentioned six

---

29 As of 1994, Chalmers Institute of Technology and Jönköping University College are also private in the sense that they are owned by foundations created by the allocation of means from the Wage Earners' Investment Funds.

30 In addition to these six institutions, a number of specialised institutions were created in

institutions were the most significant ones for the development of Swedish corpus linguistics. In addition, the Royal Institute of Technology (KTH) was important for linguistic research.

Before the 1970s, when corpus linguistics first developed in Sweden, resource allocation was highly centralized. Each year, institutions for higher education, like all other state agencies, had to submit their financial demands for the coming year to the Ministry of Education. These documents were preceded by intensive negotiations inside the universities, but sometimes also by the lobbying at the Ministry by individual professors and other university representatives for their particular interests. The following Government bill then contained very detailed prescriptions for the use of resources.[31]

In the 1970s the Swedish system of higher education institutions took a quantum leap with the creation of twelve university colleges. In the 1980s and 1990s another six university colleges were founded. In this way all Swedish counties obtained an institution of higher education (Engwall & Nybom, 2007). Most of these had the ambition to gain university status and to receive research money from the Government. So far, six of the university colleges have been upgraded to universities: Luleå in 1997, Karlstad, Örebro and Växjö in 1999, Mid Sweden University in 2005, and Malmö in 2018. However, the increase in the number of institutions also made politicians turn to the market for resource allocation. This meant

---

the nineteenth century and the early twentieth century: the Karolinska Institute (*Karolinska institutet*) medical college in Stockholm (1810); the Royal Institute of Technology (*Kungliga Tekniska Högskolan*, KTH) in Stockholm (1826), the Chalmers Institute of Technology (*Chalmers Tekniska Högskola*) in Gothenburg (1829), engineering schools; the business schools in Stockholm (1909) and Gothenburg (1923); and colleges for veterinary medicine (1914), forestry (1915) and agriculture (1932). (See further Engwall & Nybom, 2007.)

31 Needless to say, these bills did not provide everything that had been demanded in the submitted documents. They could also include surprises to the universities by providing resources for chairs they had not asked for. For instance, when Sune Carlson was inaugurated as professor at Uppsala University in 1958, he was told that a chair in business administration was not what they had asked for; the university had preferred an additional chair in astronomy. (Personal communication from Sune Carlson.)

that more resources were funnelled through research councils (see p. 33) and that grants to institutions were gradually based on performance. The increasing project financing implied that the power of individual professors over research resources was drastically reduced, unless they were members of research-funding bodies. The same was true for university leaders, who had less control over the cash flow of their institutions. With time they regained a certain modicum of power through agreements with some funding bodies that applications should be approved by the Office of the Vice-Chancellor before submission.[32]

Power relations inside institutions

Traditionally departments were run by single chairholders with some administrative support. There were also a temporary research position as *docent* (reader, associate professor), which was not tenured and could be held in principle for six years only. The possibilities of obtaining such positions were dependent on two things: (a) the budget of the faculty and (b) the grading of doctoral theses. Both were the result of professorial negotiations between and within faculties, in other words, how individual professors succeeded in defending their discipline in the creation of posts and how they managed to get support from their faculty colleagues in the thesis grading. These theses, which could be preceded by a licentiate thesis and degree, had requirements similar to the French *thèse d'état*. A top grade, decided by the faculty *in pleno*, was normally a prerequisite for an academic career (see further Engwall, 1987). Needless to say, this screening of candidates was a significant foundation for the authority structures. Another such basic element was the promotion procedures. They were based on the principles of open competition among candidates for posts that had become vacant through retirement or death of the holder as well as through the creation of new posts. The screening of candidates was done by a committee of

32 This is for instance the case with the Knut and Alice Wallenberg Foundation (Dahlberg, Hedenqvist & Sundström, 2017, p. 98).

disciplinary experts, at the time often with the chairholder as a member. The latter thus implied that the authority of the chairholders could also be extended after their retirement.

The above implies that, all in all, chairholders in the 1950s had considerable power within small departments. However, in the 1960s the situation changed considerably, as the number of students increased strongly from around 14,000 in 1946 to 25,000 in 1955 and to 69,000 in 1965 (*Statistical Yearbook of Sweden 1956*, Tables 356, 359 and 1966, Table 351). Behind this expansion were demographical factors as well as the absence of restrictions on student numbers within the faculties of Humanities, Social Sciences and Natural Sciences. As a response to this expansion of student bodies, a new position was established in the Swedish system in 1958: lecturer (*universitetslektor*), dedicated solely to teaching at the undergraduate level. In this way professorial control of university departments became reduced. This development was reinforced by a general democratization of universities in the wake of student unrest in the late 1960s. In due course, in 1977, more structured study programmes and limitations on the number of students were introduced (*Högskoleförordningen* 1977:263, kap. 5).

The creation of the lecturer position implied a need to expand doctoral programmes in order to fill the new positions. Thus, in the late 1960s Sweden introduced a four-year programme, following the American PhD model (see further Engwall, 1987). As a result, the number of completed doctoral degrees rose rapidly, especially in the early 1970s.

Another effect of the creation of the lecturer position was that chairholders in many departments abstained from being the administrative head. In this way, it sometimes happened that research priorities lost out in relation to educational and administrative priorities. In recent years as a result of an increased focus on citation counting, evaluations, rankings, etc., the balance appears to have turned in the other direction in the Swedish system.[33]

---

33 Cf. e.g. Engwall (2016), Chapter 12.

A further change, as of 1986, is the possibility for lecturers who have acquired appropriate competence to be promoted to professor after an evaluation by external experts (*Högskoleverket*, 2007). In this way the number of full professors has increased considerably.[34] At the same time the career opportunities offered to Swedish academics are still far from the United States-type tenure track system. In 2016 a government committee (SOU 2016:29) made a proposal in that direction.

External funding

The centralized resource allocation and the substantial power of chairholders over their departments can be considered a strong obstacle to innovators within various disciplines. If their professors did not approve of their preferred research orientation, the innovators could experience difficulties in their careers. External funding beyond the control of professors could therefore provide an opportunity for innovation. One early organization in this context is the Knut and Alice Wallenberg Foundation, which was founded as early as 1917 (Hoppe, Nylander & Olsson, 1993; Dahlberg, Hedenqvist & Sundström, 2017; Engwall, 2018). It was later followed by a number of other private foundations like the Wenner-Gren Foundations (1937), the Axel and Margaret Ax:son Johnson Foundation (1947), the Åke Wiberg Foundation (1954), the Torsten and Ragnar Söderberg Foundations (1960), the Sven and Dagmar Salén Foundation (1968), and the Kjell and Märta Beijer Foundation (1974).[35]

However, as early as in the 1940s the Swedish Government, inspired by initiatives in the United Kingdom and the United States, decided to create research councils in order to allocate resources to individual researchers

---

34 In recent years the practice of internal promotion has been discontinued in some of the universities, for instance the Karolinska Institute (see https://ki.se/nyheter/sa-gar-det-till-att-bli-professor-pa-ki, accessed on February 15, 2018).

35 On the Wenner-Gren Foundations, see Wallander (2002).

through project grants. This started in 1942 with one research council for technical research and another for building research. Over the following five years similar organizations were created for agricultural research, medical research, natural science research and social science research (Nybom, 1997, pp. 42–104). In 1947 the Foundation for the Humanities, which had been created by the Royal Swedish Academy of Letters, History and Antiquities in 1927, was given a similar status (Jonsson, 2003, pp. 146–149; Nybom, 1997). As will be evident below, this organization became important for the development of corpus linguistics in Sweden. In 1977, it was merged with the Research Council for the Social Sciences into the Swedish Council for Research in the Humanities and Social Science aimed at funding basic research in the areas of humanities, social science, law and theology.

The aims of the research councils were particularly to identify research needs, promote competition on a national level and to muster research resources on the international research front (Brundenius, Göransson & Ågren, 2008; Engwall & Nybom, 2007; Öhrström, 1991). Most of the research councils were subordinated to the Ministry of Education and developed into important sources of external funding for public research. As such, they constituted a complement to state block grants, which still formed the main funding source (Engwall & Nybom, 2007). The research councils supported basic research, and self-governance was their *modus operandi*. They primarily supported individual researchers or groups of researchers on the basis of their research proposals. Their organization largely corresponded to a structure based on disciplines, university departments and chairs (Skoie, 2001). The research councils were governed by scientific elites elected by peers at the universities (Bauer, 1999).

An additional significant event for the funding of the research in the humanities and the social sciences was a decision in 1964, after two years of preparations, in the Swedish Parliament to create a new free-standing research foundation to commemorate the tercentenary in 1968 of the oldest still-existing central bank in the world, *Sveriges Riksbank* (The Central Bank of Sweden). The foundation was financed through a grant of MSEK 340 from the Central Bank (*Hinc robur et securitas*, 2004, pp. 19–24). In this way

the new foundation was able to distribute twice as much as the joint budget of the research councils for the humanities and social sciences. Needless to say, this implied a significant injection of funding for the research in these areas. The foundation also played an important role for the development of corpus linguistics in Sweden.

In 1970s the research council organization was slightly restructured through mergers between some of the smaller organizations as well as the creation of a Council for Planning and Co-ordination of Research (*Forskningsrådsnämnden*, FRN) (SOU 1975:26; Premfors, 1986; Landberg, Edqvist & Svedin, 1995). In addition, resources were added to the system through the creation of research-funding bodies by various ministries and government agencies. After growing criticism of these, they were given a more research council-like character in the late 1980s (Elzinga, 1985; Gustavsson, 1989).

A more radical change in external funding occurred in the 1990s when the Parliament decided to create a number of autonomous research foundations with means from the Wage-earners' Investment Funds (*Regeringens proposition 1991/92:92*). For the humanities and the social sciences, this meant that the Bank of Sweden Tercentenary Foundation received a considerable injection of new financial resources (*Hinc robur et securitas*, 2004; Sörlin, 2005). This further reinforced the process, implying that an increasing share of state research funding was distributed on a competitive basis.

The research allocation system underwent yet another restructuring in 2001 when the basic research councils were merged into one organization, the Swedish Research Council (*Vetenskapsrådet,* VR). At the same time, funding bodies for applied research were amalgamated into three organizations addressing research on innovation (VINNOVA), sustainable development (FORMAS) and working life (FAS), respectively. In addition, the government bill (*Regeringens proposition 2000/01:3*) pointed out strategic research areas as well as the need for interdisciplinary and multidisciplinary research. Finally, there has been a tendency for the funding bodies to favour large projects to 'strong environments' or 'centres of excellence'.

The above implies that Sweden has relatively long traditions of external

funding. In the course of time, the share of this type of financing has increased. As already mentioned, and as will be shown below, both the Bank of Sweden Tercentenary Foundation and the Research Council for the Humanities have been important for corpus linguistics. VINNOVA and its predecessors have played a similar role for phonetics and related research.

Conclusions

In terms of institutional conditions for the development of corpus linguistics, we can thus conclude that professors traditionally had considerable power over their departments, although they had to negotiate with representatives of other disciplines to attain resources and support for their collaborators in their research careers. In the course of time their power was reduced as lecturers were recruited in order to handle the growing population of students. Their control over research was also diminished by the development of research councils and research foundations, through which individual researchers were able to receive grants outside the normal budgetary process. For corpus linguistics the Research Council for the Humanities and the Bank of Sweden Tercentenary Foundation were particularly significant in opening up the system, despite the fact that these bodies were controlled by scientific elites.

Developments over the past two decades have brought a tougher environment for the individual researchers in attracting research resources through increasing competition but also through a tendency of research-funding bodies to favour large grants. In a way, this has once again increased the power of established professors in relation to their younger colleagues.

CHAPTER 4. DISCIPLINARY STRUCTURES IN SWEDEN

Introduction

Language studies have long been part of academic research in Sweden. A chair in Hebrew was created at Uppsala University as early as 1605 (Isaksson & Malmberg, 2005), followed over time by professorships in other languages. This led to the creation of faculty structures with departments for various languages, but also departments of general linguistics (*allmän språkvetenskap*) in the 1960s and 1970s. In the former departments, research focused on a specific language (like English, French, German, etc.), while the latter took an interest in the comparison between languages and identification of general linguistic patterns. In the language departments, which have been described as being characterized by tradition (Enkvist et al., 1992, p. 14), philology and traditional historical linguistics dominated until the end of the Second World War. However, the time period following the Second World War brought a remarkable change and the introduction of new streams of research, particularly in phonetics, Nordic languages and English in Uppsala, Lund, Stockholm and Gothenburg (Enkvist et al., 1992).

Uppsala

According to the national evaluation in the early 1990s (Enkvist et al., 1992), language studies at Uppsala University up to the 1960s were characterized by strong conservatism. Although Uppsala was not alone in being traditional in that sense, it was more resistant to concepts like phonology and structuralism than its national counterparts in Lund, Stockholm and Gothenburg (ibid., pp. 94–96). This conservative attitude seems to have been based on a strong orientation in the 1930s towards traditional

linguistics and philology. However, there is evidence of a growing interest in general linguistics and linguistic theory in the early 1940s through a series of lectures arranged by the Linguistic Society in Uppsala (*Språkvetenskapliga sällskapet i Uppsala*, see Nyberg, 1943).[36] From the early 1950s and onwards, Uppsala developed a competence in phonetics, through the work of the Romanist Göran Hammarström (b. 1922).[37] Among his students were Björn Lindblom (b. 1934) and Sven Öhman (1936–2008), who after licentiate degrees in Uppsala defended their dissertations (Lindblom, 1968; Öhman, 1968) in Lund and at the Royal Institute of Technology, respectively. In 1965 Hammarström became the first professor in phonetics at Uppsala, but he left after a year to take up a professorship in Australia. His successor in 1969 was Sven Öhman (Enkvist et al., 1992, pp. 94–96).

One of Sven Öhman's students was Per Linell (b. 1944), who after a PhD in linguistics in 1974 was appointed to the first chair of communication studies at Linköping University in 1981. At Linköping Per Linell worked in an organization which deviated from the normal way of organizing education and research in the Swedish system. In this new university, postgraduate courses and research staff were structured in terms of multi- and interdisciplinary teams instead of traditional departments (ibid., p. 98).

Corpus linguistics was introduced in Uppsala through the work of Jan Svartvik (b. 1931). He was a doctoral student in the English department, where he presented his licentiate thesis in 1961 and the doctoral dissertation in 1966 (Svartvik, 1966). He could do so despite the fact that his two professors Erik Tengstrand (1898–1984) and Heinrich W. Donner (1904–1980) had completely different expertise: Old and Middle English philology and the nineteenth-century poet Thomas Lovell Beddoes, respectively (see Chapter 5, pp. 52–54).

---

36 In the early 1940s the founder of the Prague school, Roman Jakobson (1896–1982), spent some time in Uppsala and even published a book on children's language and aphasia (Jakobson, 1941) there.

37 However, according to Rundgren (1978, p. 98), Uppsala linguists could still in the 1950s now and then hear the statement 'modern linguistics came to a halt in Copenhagen' (our translation).

By the early 1990s Svartvik had attracted many followers. The evaluation of Swedish linguistics lists as many as 21 corpora at Uppsala University (list in alphabetical order after an extraction from Enkvist et al., 1992, pp. 103–114):

Alarm Calls (Bengt Nordberg, FUMS)[38]
Bulgarian Poetic Language (Sven Gustavsson, Department of Slavic Languages)
Conceptual Worlds of Doctors and Laymen (Ulla Melander Marttala, FUMS)
Conversational Style of Adolescents (Bengt Nordberg, FUMS)
Development of Discourse Skills in Schoolchildren (Birgitta Garme, FUMS)
Early Modern Swedish Text Bank (Mats Thelander, Department of Scandinavian Languages)
IRIS: Immigrant Voices in Sweden – Phonetic Models (Sven Öhman, Department of Linguistics)
LSP Tests in the 20th Century (Britt-Louise Gunnarsson, FUMS)
Olof Dalin: Then Swänska Argus 1732–34 (Carin Östman, FUMS)
Popular Science Corpus (Lennart Lönngren, Department of Slavic Languages)
Russian Stem Dictionary (Anna Sågvall Hein, Department of Linguistics)
Russian Text Corpus (Lennart Lönngren, Department of Slavic Languages)
Russian Word-Form Dictionary (Anna Sågvall Hein, Department of Linguistics)
Stem Dictionary for Automatic Morphological Analysis I and II (Anna Sågvall Hein in collaboration with Christian Sjögren, Gothenburg)

38 The acronym FUMS refers to a section in the Department of Nordic Languages at Uppsala University especially focusing on modern Swedish (**F**orskning och **U**tbildning i **M**odern **S**venska).

Sven Hof: Swänska språkets rätta skrifsätt (1753) (Mats Thelander, Department of Scandinavian Languages)
Swedish Language for Specific Purposes 1730–1985 (Britt-Louise Gunnarsson, FUMS)
The Child's Linguistic Identification (Olle Hammarmo, FUMS)
Uppsala Corpus of Catalan Newspaper Texts (Ingmar Söhrman, Department of Romance Languages)
Uppsala Corpus of French Newspaper Texts (Mats Forsgren, Department of Romance Languages)
Uppsala Corpus of Italian Newspaper Texts (Lars Larsson and Ingmar Söhrman, Department of Romance Languages)

These corpora show that corpus linguistics in the early 1990s had spread to many different language departments, where different linguistic aspects were studied for old as well as modern texts, both spoken language and in print.

Lund

At Lund, chairs in modern languages (English, German, Romance and Slavic languages) and comparative (Indo-European) linguistics were created around 1910. As in Uppsala, historical aspects were dominant in studies of languages (philology). As early as 1881, Lund also established a specific forum of linguistic discussion, The Philological Society (*Filologiska Sällskapet*).

Developments in Lund from the 1930s onwards took place mainly within two topics, phonetics and general linguistics. A significant actor in that context was the Romanist Bertil Malmberg (1913–1994). As a response to intra-university forces, but also political attempts to establish a chair in phonetics (mainly by the Faculty of Philosophy) at Lund University in the late 1940s, a chair was established in 1949 by a resolution of the Swedish Parliament. Bertil Malmberg was appointed as its first holder the following year. In 1959, a chair was also established in general linguistics as a result

of transforming the chair of comparative linguistics. Malmberg now left the phonetics chair and became the first holder of the chair in general linguistics. Over the years he published numerous books on linguistic topics (Malmberg, 1954; 1963; 1966; 1970; 1977).[39] He was also one of the founders of a new journal for general linguistics and comparative linguistics, *Studia Linguistica*, in 1947 (Enkvist et al., 1992, pp. 89–91). Malmberg was the supervisor of Bengt Sigurd (1928–2010), who became an important actor in Swedish linguistic research: he held chairs first in Stockholm (1970–1978) and then in Lund (1979–1993). Another doctor from Lund was Ulf Teleman (b. 1934), who applied Chomsky's ideas to modern Swedish in his dissertation (Teleman, 1969).[40]

The pioneer in corpus linguistics in Lund was Inger Rosengren (b. 1934), who after her dissertation on adjectives in Middle High German (Rosengren, 1966) launched a study of word frequencies in two German newspapers. Later on, Jan Svartvik, who was appointed professor of English at Lund, created the London-Lund corpus of spoken language (see Chapter 5, pp. 54 and 52).

In the early 1990s the evaluators listed the following eleven corpora (list in alphabetical order after an extraction from Enkvist et al., 1992, pp. 103–114):

Bruksprosa 70 (Department of Nordic Languages)
Business Letters (Inger Rosengren, Department of German)
Children's Speech Database (Department of Linguistics)
Conversation and Debate (Department of Scandinavian Languages)

---

39 A number of Malmberg's works were translated into other languages. Malmberg (1954) was published in 16 editions, the last one in 1993, just a year before his death.

40 Teleman was a student at the Department of Nordic Languages and had no supervisor, since his professors were indisposed by illness. The dissertation was a collection of published papers. (Personal communication from Ulf Teleman.) As early as 1973 Teleman was appointed to a chair of general linguistics in Roskilde 1973, where he stayed until 1982 when he returned to Lund as professor of Swedish language. He stayed on this post until his retirement in 1999 (*Vem är det 2007*).

German and Swedish Cooking Recipes (Inger Rosengren, Department of German)
Gymnasistprosa 70 (Department of Scandinavian Languages)
Interviews from Borås (Department of Scandinavian Languages)
JUBA (Lubomír Durovič and Terho Paulsson, Serbo-Croationa/Croation and Swedish).
London-Lund Corpus of Spoken English (Jan Svartvik, Department of English)
Lund Corpus of German Newspaper Texts (Inger Rosengren, Department of German)
Recurrent Word Combinations in the London-Lund Corpus (Bengt Altenberg and Mats Eeg-Olofsson, Department of English)

Again there is evidence that corpus work had spread to several departments, although less so than in Uppsala. Both spoken and printed material was included, although limited to modern works.

Stockholm

As early as the 1920s and 1930s, the then Stockholm University College set up a broad language programme with the establishment of chairs in Nordic languages (1927), English (1932), German (1929) and Romance languages (1937) (Tunberg, 1957, pp. 177–192). In the early 1950s, following initiatives by the readers (*docenter*) in Slavic languages Birger Calleman (1902–1993) and Romance languages Max Gorosch (1912–1983), two laboratories were set up: the Phonetics Research Laboratory (*Fonetiska forskningslaboratoriet*) and the Language Training Laboratory (*Fonetiska övningslaboratoriet*), the latter intended to serve the modern language departments with recorded material and listening facilities (Enkvist et al., 1992, p. 92).[41] An outcome of

41 The laboratories shared the same quarters, equipped with a sound-conditioned studio for recording and with listening booths for the students.

the work in the research laboratory was the dissertation of Claes-Christian Elert (1923–2015; Elert, 1964), who became the first professor of phonetics at Umeå University in 1969 (Enkvist et al., 1992, p. 98).

In Stockholm, the diffusion of information about new trends in linguistics mainly took place at the Linguistic Circle (*Språkcirkeln*), and the independent Research Group for Quantitative Linguistics (*KVAL-gruppen*) as well as at the Royal Institute of Technology. The Research Group for Quantitative Linguistics was initiated in 1964 by Hans Karlgren (1933–1996) and Benny Brodda (b. 1934) in order to study language with statistical and quantitative methods. The group also published a periodical at irregular intervals.

In the academic year 1966–1967 a new era was initiated at the university with the funding of a chair in general linguistics, and the establishment of a department of linguistics. Karl-Hampus Dahlstedt (1917–1996) was appointed to the chair in 1967 but left after two years for a similar chair in Umeå. He was succeeded by Bengt Sigurd (1928–2010), at the time reader at Lund University, who opened up for new research directions in syntactic studies, psycholinguistics, text linguistics and computational linguistics. In addition, two series of publications were started at this time (ibid., p. 94).

Another important activity taking place in Stockholm was the arranging of the first symposium on the description of the Swedish language (*Svenskans beskrivning*) in 1963. This event was to be followed by additional symposia in Sweden and later also in Finland, with proceedings published on a regular basis (ibid., p. 93).

In addition to Stockholm University the area had another important institution for linguistic studies: the Department of Speech Transmission at the Royal Institute of Technology (*Kungliga Tekniska Högskolan*, KTH). It was set up by Gunnar Fant (1919–2009), a pioneer in the acoustic theory of speech production (cf. e.g. Jakobson, Fant & Halle, 1961). The department housed a research laboratory, which was soon recognized internationally as a centre of excellence (Enkvist et al., 1992, p. 94).

In terms of corpus linguistics, it was the professor of Romance languages Olof Brattö (1915–2007), who started out by developing a corpus based on modern Italian novels. This was a far cry from his earlier research on proper names in Florence in the thirteenth century (cf. e.g. his dissertation, Brattö, 1953) as well as that of his predecessor Gunnar Tilander (1894–1973) who specialized in French hunting terms in the Middle Ages (cf. e.g. Tilander, 1957). For various reasons his corpus never materialized, but he inspired a follower in French through his student Gunnel Engwall (b. 1942) (see further Chapter 5, p. 55). Some twenty years after her dissertation (Engwall, 1974), at the time of the above-mentioned evaluation, the number of corpora in Stockholm had increased to eleven (listed in alphabetical order after an extraction from Enkvist et al., 1992, pp. 103–114):

KTH Speech Database (Department of Speech Communication and Music Acoustics, the Royal Institute of Technology)
Savonarola Corpus (Jane Nystedt, Department of Romance Languages)
Stockholm Bilingual and Learner Corpora (Åke Viberg, Center for Research on Bilingualism)
Stockholm Corpus of English Newspaper Texts (Magnus Ljung, Department of English)
Stockholm Corpus of French Best-Selling Novels (Gunnel Engwall, Department of Romance Languages)
Stockholm Corpus of French Economic Texts (Gunnel Engwall and Sune Stöök, Department of Romance Languages)
Stockholm Corpus of French Newspaper Texts (Gunnel Engwall and Inge Bartning, Department of Romance Languages)
Stockholm-Umeå Corpus of Modern Written Swedish (Gunnel Källgren, Department of Linguistics with Eva Ejerhed in Umeå)
Swedish TEFL Corpus (Magnus Ljung, Department of English)
Swedish-French Bilingual Children Database (Department of Linguistics in collaborations with Department of Romance Languages at Lund University)
The FIDUS Corpus (Erling Wande, Department of Finnish)

The number of corpora was thus the same as in Lund and, as in Uppsala and Lund, with both spoken and printed material. Most of the corpora were found in the Department of Romance Languages, while others had been created for English, Finnish and Swedish.

Gothenburg

From its foundation in 1891 the Gothenburg University College had a number of chairs in language studies, including comparative linguistics. Among the first professors was Gustaf Stern (1882–1948) in the area of the English language, known for his work in historical semantics. Another early professor was Bernhard Karlgren (1889–1978), a pioneer in the study of the history of Chinese languages. However, it was at the Department of Nordic Languages that corpus linguistics developed in Gothenburg through Sture Allén (b. 1928), a student of the philologist Ture Johannisson (1903–1990) (see Chapter 5, p. 49). In parallel, modern linguistics, and particularly the ideas of Chomsky, was introduced in Gothenburg through Alvar Ellegård (1919–2008), who was appointed to the chair in English in 1962. His research interest covered a number of areas including studies of English historical syntax, transformational grammar and contrastive and applied studies (cf. e.g. Ellegård, 1953; 1962; 1978). His popular writings introduced modern linguistics to a broader audience and contributed to an increased consciousness of several scholars in Sweden. His work also transformed the University of Gothenburg into a strong centre of English studies (Enkvist et al., 1992, pp. 96–97).

As will be evident in Chapter 5 (pp. 49–52), Gothenburg became very central in terms of the development of corpora over the years, particularly for Swedish. An extraction from Enkvist et al. (ibid., pp. 103–114) yields the following twelve corpora in alphabetical order at the University of Gothenburg and Chalmers Institute of Technology in the early 1990s:

Corpus of American Collocations (Göran Kjellmer, Department of English)
CTH Speech Database (Department of Information Theory, Chalmers Technical University)
Gothenburg Corpora of Spanish Texts: PEE77 and ONE71 (David Mighetto and Per Rosengren, Department of Romance Languages)
GREVOC: Greek Vocabulary (Bo-Lennart Eklund, Department of Classical Languages)
Legal Language (Department of Computational Linguistics)
Novels 76 (Department of Computational Linguistics)
Novels 80 (Department of Computational Linguistics)
Parliamentary Debates (Department of Computational Linguistics)
POLSVE (Roman Laskowski, Department of Slavic Languages).
Press 65 (Department of Computational Linguistics)
Press 76 (Department of Computational Linguistics)
Press 87 (Department of Computational Linguistics)

It is evident that most of these corpora covered Swedish, which is natural in relation to the research programme pursued by Sture Allén. In relation to the other universities Gothenburg is the only one with a corpus of a classical language.

Conclusions

Language studies have long been an important part of the Swedish academic system. Traditionally these studies have been organized within departments specializing in individual languages. From the 1950s and onwards there was a development of general linguistics, which was strongly related to phonetics, particularly in Lund and Stockholm. This has led to the creation of chairs in general linguistics and phonetics. In the last decade there has also been a tendency to merge language departments to larger units such as departments of modern languages.

It is particularly worth noting that computational linguistics has been integrated into language departments, contrasting to the international situation where computational linguistics or natural language processing (NLP) is often found as a specific branch within departments of electrical engineering or computer science. As such, the Swedish approach has been more linguistically and grammatically oriented (Enkvist et al., 1992, p. 20).

In terms of corpus linguistics, pioneers appeared at all four of the older universities: Uppsala, Lund, Stockholm and Gothenburg. These innovators were not working in departments of general linguistics, but in specific language departments: English, German, French and Swedish. In this development the old professors seem to have been supportive rather than critical, which may be explained by the fact that the collection of examples and the use of corpora have a long tradition in Swedish language research. Criticism of corpus linguistics was instead more voiced by representatives of the newly created discipline of general linguistics, particularly those who adhered to the ideas of Noam Chomsky and his transformational grammar. However, by the early 1990s the use of corpora appears to have become a widely used approach in language research. Enkvist et al. (1992) thus in total identified as many as 55 corpora in Uppsala, Lund, Stockholm and Gothenburg. Including corpora in Linköping (Linköping Discourse Corpus, and Man-Machine Dialogues) as well as in Umeå (The Structure and Verbal Skills among Pupils, Umeå Speech Database, and Umeå Corpus of French Newspaper Texts) this figure increases to 60. A similar count for the present day is likely to produce a much higher figure.

CHAPTER 5. A FIRST GENERATION OF SWEDISH INNOVATORS

Introduction

In the previous chapter, it was pointed out that language studies were traditionally carried out in departments with a specialization in specific languages, such as the classical languages, modern foreign languages and Nordic languages. It was not until the 1960s that chairs and department of general linguistics were created. These were thus fewer than the traditional language departments and often had a different orientation (read Chomsky). It is therefore quite natural that the development of corpus linguistics occurred in the language departments. This chapter will present four Swedish pioneers in the field: Sture Allén in Gothenburg for Swedish, Jan Svartvik in Uppsala, London and Lund for English, Inger Rosengren in Lund for German, and Gunnel Engwall in Stockholm for French.

The pioneer for Swedish: Sture Allén in Gothenburg

The main early actor to create a corpus of Swedish language to be studied by means of computers is Sture Allén (b. 1928).[42] He had technical inspiration already at home in Gothenburg, since his father was an engineer working for a company that constructed and sold safes. The orientation towards natural sciences continued in school, where a number of his friends aimed

42 In addition to the referred sources this section is based on an interview with Sture Allén, by Lars Engwall on November 17, 2011.

at studies at Chalmers Institute of Technology. Allén also had a strong interest in the humanities, and in addition to natural science studies took languages and philosophy at school. After leaving school he decided to study Nordic languages at the University of Gothenburg and prepared himself for these studies during his compulsory military service by acquiring the secondary school qualifications in Latin. The studies at the time in Nordic Languages at the University of Gothenburg were strongly directed towards language history and old epochs. Allén has pointed out in an interview that his language studies thus included the reading of the bible of Wulfila in Gothic, the whole Edda in Old Islandic and Old Swedish laws.[43]

In addition to Nordic languages, Allén studied English, Literature and Psychology, whereupon he was taken on as an assistant in the Department of Nordic Languages and started his work for a Licentiate (at the time almost equivalent to a PhD). Together with his supervisor Ture Johannisson (1903–1990) he decided to focus on seventeenth-century language and particularly the letters of Johan Ekeblad to his brother, father and other persons in the mid-seventeenth century. This research resulted in a dissertation (Allén, 1965) consisting of two parts: one was a commentated edition of the letters of Ekeblad, the other a presentation of a method for the analysis of the text (*Grafematisk analys som grundval för textedering* ('Graphemic analysis as a basis for text editing'). This in turn inspired him to contact the Computer Institute at Chalmers and to learn programming in machine code (Peralta, 2008, p. 3). His interest in computer use had also been manifested a year before his thesis defence in a book review in one of the Gothenburg dailies (Allén, 1964) under the title 'Ordforskaren och datamaskinen' ('"The linguist and the computer").

After his dissertation Allén set up a research group financed by the Bank of Sweden Tercentenary Foundation and the Council for Research in the Humanities to study modern Swedish by means of computers. The corpus consisted of one million words from morning newspapers in the three

43 The paragraph draws upon Peralta (2008).

largest Swedish cities (Stockholm, Gothenburg and Malmö). A first output of the programme was the dictionary *Nusvensk frekvensordbok* ('Frequency Dictionary of Present-Day Swedish') in four volumes (Allén, 1970–1980). During the same period the group also published a more condensed frequency dictionary *Tiotusen i topp* ('Top Ten Thousand', Allén, 1972) and a dictionary of homographs, *Olika lika ord* ('Different Similar Words', Berg, 1978). The research also led to the foundation of the Language Bank (*Språkbanken*; see http://spraakbanken.gu.se/), which was given the task of collecting, storing, processing and providing Swedish texts that could be read electronically. It was established in 1975 as a national centre of computational lexicography, and as such, it also became the first department of computational linguistics in Sweden. Through this centre, corpora users have been able to access linguistic and statistical data about a diverse range of Swedish texts since the 1970s. On the basis of the material in *Språkbanken*, in the mid-1980s the group published *Svensk ordbok* ('Swedish Dictionary', Abelin & Allén, 1986), which developed into *Nationalencyklopedins ordbok* ('Dictionary of the Swedish National Encyclopaedia', 1995–96).[44]

As mentioned, Allén's research was supported by the Bank of Sweden Tercentenary Foundation and the Council for Research in the Humanities. In 1970 he obtained a special research position at the Research Council for the Humanities, and in 1972 he was appointed Professor of Computational Linguistics (*språklig databehandling*) at the Research Council. In 1979 this professorship was taken over by the University of Gothenburg, where Allén was a professor until his retirement in 1993. In the meantime he was elected one of the eighteen members of the Swedish Academy in 1980, where he was Permanent Secretary between 1986 and 1999. Before taking up the latter position he was Deputy Vice-Chancellor of the University of Gothenburg 1980–86.[45]

---

44 The text is based on the presentation of Sture Allén at the website of the Swedish Academy, http://www.svenskaakademien.se/svenska-akademien/de-aderton/stol-nr-3-sture-allen, accessed on July 29, 2017. For a collection of Allén's published papers, see Allén (1999).

45 This paragraph is based on *Vem är det 1997*.

There is no doubt that Sture Allén was an academic entrepreneur, who succeeded in introducing a new approach to the study of Swedish. However, it is also worth noting that the timing of his innovation was fortuitous. One significant financier of his research was the Bank of Sweden Tercentenary Foundation (cf. Chapter 3, pp. 34–35), which in the mid-sixties was searching for large innovative projects in the humanities and social sciences. It is also noteworthy that he received special treatment from the Research Council for the Humanities through positions as a researcher and professor.

The pioneer for English: Jan Svartvik in Uppsala, London and Lund

Among students of English, the pioneer in Sweden in terms of corpus linguistics is Jan Svartvik (b. 1931).[46] He first came into the field in 1959 after having found in the English Department Library at Uppsala University a paper by Randolph Quirk entitled 'Relative Clauses in Educated Spoken English' (Quirk, 1959) in the journal *English Studies*. This inspired him to apply for a grant from the British Council for a year of study at University of Durham, where Quirk was working at the time. After his return to Uppsala University he had his Licentiate dissertation accepted. Only a week after that he received an invitation from Quirk, who had moved to the University College London, to work as his research assistant. As a result, Svartvik spent the period 1961–1965 as first research assistant and later deputy director of the above-mentioned project Survey of English Usage (SEU) run by Randolph Quirk. This work led to his doctoral dissertation at Uppsala University entitled *On Voice in the English Verb* (Svartvik, 1966). Together with Quirk he also published in the same year *Investigating Linguistic Acceptability* (Quirk & Svartvik, 1966). The continued collaboration with Quirk resulted in *A Grammar of Contemporary English* (Quirk & Svartvik, 1972).

---

46 In addition to the sources listed, the text is based on an interview with Jan Svartvik, by Lars Engwall on December 2, 2011. See also Svartvik (2005).

In 1975 Svartvik took the initiative for a sister project of the London Survey: Survey of Spoken English (SSE). While SEU consisted of both written and spoken English, SSE focused on spoken English. Together the two surveys resulted in a corpus of one million words from 100 written materials and 100 spoken materials. The corpus was presented in *A Corpus of English Conversation* (Svartvik & Quirk, 1980) as well as in *The London-Lund Corpus of Spoken English: Description and Research* (Svartvik, 1990). The latter work also provided research results from the use of the corpus. Between these two publications the London-Lund team also published *A Comprehensive Grammar of the English Language* (Quirk et al., 1985).

After a guest professorship at Brown University with Kučera and Francis (cf. Chapter 2), Svartvik was appointed to the Chair of English at Lund University in 1970 and remained in this position until his retirement in 1996. Over the years he has garnered considerable recognition for his work. He is an elected member of a number of learned societies, among them the Royal Swedish Academy of Letters, History and Antiquities (1981), the Academia Europaea (1989), the Royal Swedish Academy of Sciences (1990), and the New York Academy of Sciences (1994). Svartvik has also been awarded honorary degrees from the University of Bergen, Masarykovy University, Brno, and the University of Helsinki. Furthermore he has been the chairman of the *Association internationale de linguistique appliquée* (1981–1984) and member of the Swedish Research Council for the Humanities and the Social Sciences (1980–1986).[47]

Jan Svartvik constitutes a case of early adoption of an international development. His wish to further pursue the ideas of Randolph Quirk does not seem to have met with any resistance from his professors (cf. Chapter 4, p. 38), who rather supported his plans to go to the United Kingdom. He joined Quirk at the right time and could in this way be involved in a significant project on spoken English. He thereby became closely connected

47 The information in this section is based on *Vem är det 2007*. See also his Festschrift (Aijmer & Altenberg, 1991).

to a group of important British colleagues and could also later on develop considerable corpora with Swedish funding.

The pioneer for German: Inger Rosengren in Lund

For the development of a German corpus in Sweden Inger Rosengren (b. 1934) at Lund University was the pioneer. Even in her thesis, *Semantische Strukturen: eine quantitative Distributionsanalyse einiger mittelhochdeutscher Adjektive* (Rosengren, 1966), defended at Lund, she had used quantitative methods for an analysis of adjectives in Middle High German (*Mittelhochdeutsch*, i.e. German in the period 1050 and 1350). Her external examiner at the thesis defence was Sture Allén, who is likely to have inspired her to move in the direction of corpus linguistics. Her decision to do so was facilitated by the fact that her dissertation was graded as qualifying for the title of *docent* (reader, associate professor, see Chapter 3, p. 31).[48] Her project on corpus linguistics was supported by the Swedish Research Councils for the Humanities and for the Social Sciences, as well as two foundations (Carl-Bertel Nathhorsts vetenskapliga stiftelse and Längmanska kulturfonden).[49] Like Sture Allén, Inger Rosengren turned to newspaper texts.[50] In so doing, she could take advantage of the then modern technology in newspaper production: she managed to get access to the six-channel magnetic tapes that had been used for the type-setting of the two German newspapers *Die Welt* and *Süddeutsche Zeitung*. From these she excluded certain categories and

48 As mentioned in Chapter 3, in the old system Swedish doctoral dissertations were graded. The top grades qualified their authors the title of *docent*, which was virtually a prerequisite for an academic career.

49 See Rosengren (1972, p. VI).

50 Interestingly enough, her husband, Karl Erik, was a media researcher who had started his academic career at the Department of Literature Department at Lund University. He left this department for the Department of Sociology as his supervisor had not approved his idea to use quantitative methods in his doctoral dissertation (Windahl, 2013).

sampled texts for the period November 1, 1966 to October 30, 1967, ending up with a corpus of close to three million running words (2,476,560 for *Die Welt* and 500,334 for *Süddeutsche Zeitung* (Rosengren, 1972, p. XXIV). The data processing was undertaken at the computer centre Medicindata in Gothenburg (a *Saab D 21*) using adaptations of programs that had been developed by the Allén research group (ibid., p. V). The project produced frequencies of German words, which were published in two volumes in the 1970s (Rosengren, 1972; 1977).

Inger Rosengren no doubt took advantage of the technological development. She appears to have been inspired by Sture Allén and his approach to use the magnetic tapes from newspapers in order to create her corpus. Corpus linguistics did not continue to be Rosengren's main research interest, however. Her corpus work qualified her for a chair in Germanic languages at Lund University in 1971. At the time she turned to more general linguistics, publishing particularly in the field of pragmatics (cf. e.g. Rosengren, 1981; 1984; 1986).

The pioneer for French: Gunnel Engwall in Stockholm

In terms of French corpus studies in Sweden Gunnel Engwall (b. 1942) is the pioneer.[51] After her master's degree in Latin and French she embarked upon doctoral studies with Professor Olof Brattö (1915–2007) as supervisor. Brattö was an expert in Italian and worked at the time on word frequencies in Italian, probably inspired by Antonio Zampolli (cf. Chapter 2).[52] He suggested that Gunnel Engwall should undertake similar studies for

51 In addition to the sources listed, this section is based on an interview with Gunnel Engwall, by Tina Hedmo on April 19, 2011.

52 Like Zampolli, Gunnel Engwall participated in one of the summer schools (in 1968 in Besançon) arranged by Bernard Quemada (see Chapter 2, pp. 19–20). She was even advised by Quemada to study the history of French-Swedish dictionaries for her thesis.

French.[53] He offered an assistantship in the department to do the work, an offer which was difficult to resist for a relatively fresh doctoral student without research funding. However, although the work was paid, it also involved issues to solve with the professor regarding (1) data processing, (2) sampling, and (3) the size of the corpus.

In terms of data processing the professor, probably inspired by old techniques among linguists, punched just one word per card. Gunnel Engwall came to challenge this as she learnt about the American program, KWIC (KeyWords in Context), which was used for the analyses of book titles in the library of the College of Forestry. This meant that most of the 80 columns of the computer cards could be filled with text, except for the last columns, which were used for references. Following considerable deliberation, the use of this program was accepted by the professor, who then restarted his own work, now also filling the punched cards. However, he was nevertheless still convinced that the best way to proceed was to put the results from the data processing on traditional type-written cards with one word with their frequency on each to be stored in a filing cabinet. Despite having adapted to the modern world by using computers, the professor was thus still caught up in old technology.

Regarding sampling, inspired by conversations with a social scientist, Gunnel Engwall, after discussions with her professor, turned to a more systematic approach than her supervisor advocated. In order to cover important French novels in the 1960s she used the best-selling lists published from the two French literary magazines, *Les Nouvelles littéraires* and *Le Figaro littéraire* for the period 1962–1968. Together, these lists contained 400 distinct titles, some one hundred of which could be eliminated on the basis on two criteria: (1) The authors were not born in France, and (2) The novel was not set in France after 1945. This left 161 titles for the final selection. These were sorted by the year of birth of their authors, and from

53 It should be mentioned that corpus studies also had an educational application, since lists of word frequencies were used as a basis for vocabulary tests.

the resulting list 25 novels were selected by choosing the youngest and not permitting more than one novel per author.

In this way the sample came to include works like *Les Choses* by Georges Perec, *Le Déluge* by the 2008 Nobel Laureate Jean-Marie Gustave Le Clézio, *La Chamade* by Françoise Sagan and *Élise ou la vraie vie* by Claire Etcherelli. From all the 25 novels, 20,000 words were selected from ten strata with random entries (cf. Engwall, 1994a, pp. 60–64). This implied that the corpus was determined to be half a million words. Needless to say, it was a very demanding task to handle a corpus of this size with the technology of the time, not least all the proofreading of the punched cards. In addition, the funding for data processing by the main frame computers of the time was restricted and required permanent applications for funds.

As Gunnel Engwall became a part of the network of corpus linguists and presented her dissertation project, older colleagues, like Sture Allén and Inger Rosengren, objected that the collection and analysis of such a large corpus was too much for a doctoral thesis. This issue was solved in 1971, when the professor resigned and was succeeded by Gustaf Holmér (1921–2004), an expert on medieval hunting terms. Despite his limited knowledge of computational linguistics he was supportive of his doctoral student's project. As he realized that the material was too extensive for a thesis, he asked her in 1973 to find some way to finish her dissertation. At the time, the whole material was processed on the word level, which permitted statistical tests and comparisons with English and Swedish corpora. However, the relating of all inflected forms of a word to their main word, the lemmatization, had to be limited, since this was a very time-consuming work without any of the data programs that exist today, even with assistance. For the dissertation the professor and Gunnel Engwall therefore decided to stop lemmatization after 10 of the 25 novels.

The dissertation *Fréquence et distribution du vocabulaire dans un choix de roman français* (Engwall, 1974), published by the Stockholm linguistic research group Skriptor, was defended in the spring of 1974 and paved the

way for a post-doctoral position (*forskarassistent*).[54] The latter was a position for six years with 75 per cent of the time devoted to research. During this employment Gunnel Engwall finished the work with all the 25 novels and could present the results in a frequency dictionary *Vocabulaire du roman français (1962–1968): dictionnaire des frequencies* (Engwall, 1984) published in the series *Data linguistica* edited by Sture Allén (cf. also Engwall, 1978; 1995; 1996). The material was then included in the French corpus library INaLF (cf. Chapter 2) and was used for phonetic studies by the Department of Speech, Music and Hearing at the Royal Institute of Technology. Gunnel Engwall's work in corpus linguistics also brought her into an international network, manifested by her board membership for the years 1988–1994 in the Association for Literary and Linguistic Computing (ALLC).

During her work Gunnel Engwall felt that many linguists, who were much inspired by Noam Chomsky, were highly negative towards the production of corpora. However, in the 1990s attitudes changed, and her feeling is now that presently almost all linguists use corpora. In relation to this change it is also important to remember that technological developments have made it much easier to compile corpora. In addition, corpus use is much less time-consuming and also easier than corpus compilation, not least thanks to the development of various computer programs.

In the 1980s Gunnel Engwall and her colleague Inge Bartning, as mentioned in Chapter 4, developed another French corpus, this time using French newspaper texts.[55] They developed a corpus COSTO (COrpus of STOckholm) which contained one million running words from the Paris newspaper *Le Monde* and the French weekly *L'Express*. Texts were selected through a sampling procedure for the period March 1987 to February

54 Key persons at Skriptor at the time were Hans Karlgren and Benny Brodda, mentioned in Chapter 4 (p. 43).

55 See further Engwall & Bartning (1989) and Danell (1990). Bartning later turned to second-language acquisition and has also developed corpora in that research (see http://www.su.se/profiles/bartn-1.195116, accessed on July 29, 2017).

1988 inclusive.[56] For this study magnet tapes from the publications could be used, which of course made data collecting and processing much easier than punching the entire corpus on cards. At a later stage a colleague of Gunnel Engwall, Mats Forsgren, along with Françoise Sullet-Nylander and Malin Roitman, continued the studies of language of media by turning to radio and TV. In so doing, they used the corpus FPM (*le Français Parlé des Médias*) developed with colleagues at Uppsala University (see Forsgren, 2002). It consists of 50 hours of TV material (news, debates, talk shows, etc.). The group has also studied a corpus based on the five televised debates between the two principal contenders in the French presidential elections of 1974, 1981, 1988, 1995 and 2007.

Gunnel Engwall later redirected her research interest towards Strindberg as a French author (cf. e.g. Engwall, 1980; 1990; 1994b; 1998; 2009).[57] She also got involved in university administration, first as Head of Department (1988–1994), then Pro Vice-Chancellor (1994–2003) and Acting Vice-Chancellor (2003–2004). In addition, she has been a member of various bodies for research financing and was the President of the Royal Swedish Academy of Letters, History and Antiquities from 2006 to 2013.

Gunnel Engwall is an example of a corpus builder who came into the field early, facing all the technical complications of the time. The latter were particularly associated with the specific diacritical marks used in French and the limited knowledge about corpus studies in her department. In this situation, links to French as well as Swedish colleagues, particularly Sture Allén, were highly important.

---

56 The corresponding corpora for Belgium and Switzerland included texts from *Le Soir* and *La Libre Belgique* and from *La Tribune de Genève*, respectively (Engwall, 1994a, p. 67, note 21).

57 In the last two publications corpus linguistics is used for an analysis of Strindberg's language.

Conclusions

In relation to the model presented in Chapter 1 the four cases presented above do not provide much evidence of strong resistance against corpus linguistics from the established professors. Despite the fact that most of them were oriented towards traditional language studies, they were open to the innovation of corpus linguistics. The negative views came rather from another camp, which was then under establishment: general linguistics and the ideas of Chomsky.

It is also evident that the addition of external funding was important for the course of events. This is particularly the case for Sture Allén, who got support from both the Bank of Sweden Tercentenary Foundation and the Research Council for the Humanities. Apparently, even there the established professors appear to have been open-minded.

If Allén is a good example of the significance of external funding, the case of Jan Svartvik provides the corresponding evidence for the importance of links to international developments. Svartvik's contacts with Randolph Quirk and his group were instrumental for his own research as well as for the following efforts in Sweden in corpus linguistics. In all four cases, it is of course extremely important to bear in mind the technological development. Particularly the early works of Sture Allén and Gunnel Engwall show the difficulties associated with corpus building at the time. These difficulties had been reduced considerably for the second generation of corpus linguists, which will be the topic for the following two chapters.

CHAPTER 6. A SECOND GENERATION DEALING WITH WRITTEN LANGUAGE

Introduction

The first generation of innovators were born in the 1920s, the 1930s and the early 1940s. After them came a second generation who followed the trail-blazers. They were mainly born after the Second World War and defended their dissertations in the late 1970s and after. This chapter will illustrate with two examples how the work with corpus linguistics was continued for written Swedish and English in Swedish institutions. In terms of Swedish, the work of Sture Allén continued at the University of Gothenburg through Lars Borin, a PhD from Uppsala University, presently the Director of *Språkbanken*. As for English the main researcher in Sweden is nowadays Merja Kytö at Uppsala University, who started her academic career in Finland, at the University of Helsinki. Both Borin and Kytö defended their doctoral dissertations in 1991.

From Slavic languages to *Språkbanken*: Lars Borin in Uppsala and Gothenburg

Lars Borin (b. 1957) started his doctoral education in the area of Slavic languages in the early 1980s while also working at UCDL, a unit for computational linguistics at the Uppsala University Data Centre (UDAC).[58] In 1990, this centre was moved to the Department of Linguistics, in which

58 This section is based on an interview with Lars Borin, by Tina Hedmo on August 8, 2011, with additional information from Lars Borin in April 2018.

computational linguistics became a new disciplinary area with Anna Sågvall Hein (b. 1941), the director of UCDL and one of Borin's supervisors, as the first chairholder. As a result, shortly before his thesis defence, with full consensus between the two professors involved, Borin changed from Slavic languages to computational linguistics.[59] He was therefore, with the thesis *The Automatic Induction of Morphological Regularities* (Borin, 1991), the first person to receive a PhD in computational linguistics at Uppsala University.

After his dissertation, Borin stayed at the Department of Linguistics at Uppsala for many years, first as a research fellow, and later as a senior lecturer. An important task during this period was the establishment of and teaching in a new programme in language technology at the undergraduate level.[60] In addition, he became involved in two research projects. One of these was the Uppsala Learning Lab, a project focusing on IT and learning, funded by the Knut and Alice Wallenberg Foundation linking Stanford University with Uppsala University, the Karolinska Institute and the Royal Institute of Technology.[61]

In contrast to the Uppsala Learning Lab project, the second project was corpus-oriented. It had been initiated by Borin's supervisor Anna Sågvall Hein. It was supported by the Bank of Sweden Tercentenary Foundation and was a collaboration with researchers at Stockholm University. This project focused on machine translation and interpretation of parallel corpora and led to a number of publications (e.g. Olsson & Borin, 2000).

In 2002, after a brief interlude at Stockholm University as head of the computational linguistics unit at the Department of Linguistics, Lars Borin was appointed to a chair in natural language processing in the Department

---

59 As a matter of fact, he published 'Is Hungarian a Case Language?' (Borin, 1986), which is an early contribution to corpus linguistics before his dissertation.

60 A corresponding programme had been created ten years earlier at the University of Gothenburg (see Chapter 7, pp. 76–77).

61 Later the Leibniz University Hannover was affiliated as well as Lund University. The project is still ongoing, with the support of the Knut and Alice Wallenberg Foundation, under the label Wallenberg Global Learning Lab. The host for the project is Lund University. The project mainly attracts engineering students.

of Swedish at the University of Gothenburg. This meant that he became more focused on language technology and the development of tools to facilitate the use of corpora. Basically, his research in Gothenburg has focused on three areas. The first deals with computational lexical resources, which are used above all for linguistic annotation of text corpora. The second area aims at developing tools or software for language data, while the third is the provision of digital resources for minority languages or low-resource languages, namely, those lacking resources such as corpora and lexicons.

Borin's appointment to the Gothenburg chair also implied that he became the head of *Språkbanken* ('The Swedish Language Bank'), created by Sture Allén (cf. Chapter 5, p. 51), providing access to vast text corpora for researchers in Swedish and the Nordic languages. A related activity is the website *Litteraturbanken* ('The Swedish Literature Bank'), a collaboration involving *Språkbanken*, the Swedish Academy, the National Library of Sweden, the Royal Swedish Academy of Letters, History and Antiquities, the Swedish Society for Belles Lettres, and the Society of Swedish Literature in Finland, in order to make classical Swedish literature available on the Internet.[62] Furthermore, *Språkbanken* is the national coordinator for SWE-CLARIN, the Swedish node of the European Union initiative 'Common Language Resources and Technology Infrastructure', which aims[63]

> to create an eResearch infrastructure that makes language resources (annotated audio and video recordings, text collections and corpora, lexical resources, ontologies, etc.) and tools based on language resources and language technology (speech recognizers, lemmatizers, parsers, summarizers, information extraction and text mining systems, etc.) available and readily usable to scholars of all disciplines, in particular the humanities and social sciences.

---

62 The chairman of the board is Gunnel Engwall, who developed a corpus of French modern novels in the 1960s (see Chapter 5, pp. 55–59).

63 https://spraakbanken.gu.se/swe/forskning/infrastruktur/swe-clarin, accessed on February 19, 2018.

In the wake of a national graduate school in language technology, research groups at the University of Gothenburg with counterparts at Chalmers Institute of Technology, also in Gothenburg, initiated an informal research collaboration called the Centre for Language Technology. During the years 2011–2016 the centre was a formal unit and a specific administrative unit at the University of Gothenburg, financially supported by means of internal strategic resources of the Vice-Chancellor.[64] In addition, the Centre was successful in attracting external funding, in the form of grants from the European Union and the Swedish Research Council. In relation to the latter, the Centre has particularly benefitted from the more recent strategic efforts of the Research Council in the area of research infrastructure.[65] This has been particularly advantageous, since funding of corpus building had been difficult previously.

The national graduate school also opened up for increased Nordic research collaboration including the Baltic countries and in some contexts also the area around Saint Petersburg in Russia. Among other things, Borin and *Språkbanken* have been involved in a European Union project coordinated by a language technology company in Riga, aiming at constructing a Nordic infrastructure for language technology including corpora. The project included one partner for each country in Scandinavia and the Baltic countries.

Starting in 2018, with the help of a large research infrastructure grant (MSEK 105 for the years 2018–2024) from the Swedish Research Council, matched by an equal contribution from the ten partner institutions involved (universities and public authorities), *Språkbanken* is establishing a national research infrastructure in support of research based on language data, with Lars Borin as director. The remit of the new national *Språkbanken* will be not only text corpora and language tools for working with text,

---

64 See *Strategic Plan for the Development of Research in Language Technology at the University of Gothenburg* (2009).

65 The Swedish Research Council has a council for the financing of research infrastructure (*Rådet för forskningens infrastrukturer*, RFI).

but also speech databases and tools for working with speech, as well as an infrastructure for the so-called digital humanities and social sciences.

Borin's case demonstrates how the infrastructure in terms of computer facilities and human expertise has been important for the development of corpus linguistics. It is evident that the establishment of UCDL at the Uppsala University Data Centre (UDAC) greatly contributed to the development of computational linguistics in Uppsala. It also interesting to note that Lars Borin started out in Slavic languages and defended his thesis in computational linguistics. Again, we can note the role of the Bank of Sweden Tercentenary Foundation as a funder of corpus research. Borin's case also shows how the legacy of the pioneer Sture Allén has been preserved and developed by his successors at what must be considered the Mecca of Swedish corpora: *Språkbanken* in Gothenburg.

From Old English to an international key role: Merja Kytö from Helsinki

Merja Kytö (b. 1953) has an academic background in English linguistics at the University of Helsinki, Finland, where she started her licentiate studies in the early 1980s.[66] After a short time as visiting fellow at Yale University, where she collected early American English texts from the New England area (1620–1720), she was invited to join the Helsinki Corpus project initiated by Matti Rissanen (1937–2018).[67] It was the first stratified computerized collection of English historical texts covering the period from Old English to the early 1700s representing various language-use settings, including statutes, religious treatises, handbooks, diaries, letters, fiction

66 This section is mainly based on an interview with Merja Kytö, by Tina Hedmo on November 9, 2011, with updated information provided by Kytö in April 2016.

67 Matti Rissanen had a background in studies of Old and Early Middle English (see e.g. his dissertation, Rissanen, 1967).

and trial proceedings. The project was funded by the Academy of Finland (the Finnish Research Council) and the University of Helsinki. The project ran between 1983 and the early 1990s, and one of its visible results was Merja Kytö's doctoral thesis (1991), *Variation and Diachrony, with Early American English in Focus: Studies on CAN/MAY and SHALL/WILL*.

Kytö funded her PhD-project by working as a research assistant and the secretary of the project (1983–1991). After using traditional cards for the data collection during the first half of her work, mid-way she was able to turn to corpus linguistic methodology, thereby taking advantage of the expertise at the university computing centre. However, even these experts had to be convinced of the relevance of the project in relation to computational techniques. Kytö therefore had to learn the basic methodology from colleagues as well as by attending basic courses in corpus linguistic techniques and software.[68] A valuable source of inspiration was her exchanges with the Dictionary of Old English Corpus project at the University of Toronto.

The project generated continuing funding from the Academy of Finland, which led to the publication of a number of volumes (e.g. Meurman-Solin, 1993; Rissanen, Kytö & Palander-Collin, 1993; Rissanen, Kytö & Heikkonen, 1997a and 1997b). In the mid-1990s, when the compilation project had ended and the corpus had been published, the Department of English at the University of Helsinki was granted funding for 12 years (2 x 6 years) as a centre of excellence. This in turn opened up for cooperation with researchers in the United States, the United Kingdom and Germany. For Kytö personally, the international contacts led to an appointment as secretary of ICAME (International Computer Archive of Modern and Medieval English) in the mid-1990s and as co-editor of *ICAME Journal*.[69]

It should be noted that in the beginning the Helsinki Corpus project was met with scepticism, particularly from the literary scholars at the Department of English. However, with time it has become well received.

---

68 As a result she could produce a manual for the project (Kytö, 1996).

69 See http://icame.uib.no/; Facchinetti, 2007; Renouf & Enouf, 2009, and below Chapter 8, pp. 90–93.

As a matter of fact, the Helsinki Corpus project radically changed English historical linguistics, and it can even be claimed that it saved this discipline from fading away. The Helsinki Corpus project also paved the way for more resource-demanding research among scholars in the humanities, including computers, researchers, doctoral students etc. However, other language departments in Finland were much slower to adopt corpus linguistic methodology.

When the Helsinki Corpus project had ended, Kytö was employed at the University of Tampere first as Senior Lecturer in English Philology in 1993 and then in 1994 as Associate Professor of American English Language and Literature. Since these positions provided limited research conditions, Kytö applied for chairs at Swedish universities and was in 1995 appointed to the Chair of English language at Uppsala. Although now based in Sweden, Kytö has stayed in contact with her Finnish colleagues and has collaborated with a number of them in Helsinki, Tampere and Turku.

In terms of funding, Kytö has received two grants from the Swedish Research Council, and one from the Bank of Sweden Tercentenary Foundation. She received her first Research Council grant in the late 1990s for a corpus study of speech-related texts from 1560 to 1760 drawn from trial proceedings, witness depositions, drama texts, fiction and didactic works. The project, which was carried out in collaboration with Jonathan Culpeper (b. 1966) at Lancaster University, with the aim of exploring past spoken interaction. The idea was that written records containing specimens of speech-related language could be used to collect linguistic evidence. Again, the approach was initially met with scepticism. However, today, such voices are seldom heard. Historical pragmatics, as the research framework is called, has even become one of the most popular areas in historical corpus linguistics.

Kytö's second Research Council project concerned an electronic edition of early witness depositions from criminal and ecclesiastical courts located in different parts of England. Here Kytö and two of her Uppsala PhDs, Peter J. Grund (b. 1975, now at the University of Kansas) and Terry Walker

(b. 1961, now at Mid Sweden University) transcribed some 280,000 words and presented them in a carefully arranged collection that at the same time could serve as a stratified corpus.[70] The project resulted in a book (Kytö, Grund & Walker, 2011) accompanied by a CD-ROM containing the corpus and a customized search engine, making the materials easily accessible to users.

Kytö's third corpus linguistic project is funded by the Bank of Sweden Tercentenary Foundation (2016–2018). It investigates the use of two groups of intensifiers, namely amplifiers scaling upwards (e.g. 'terribly', 'most') and downtoners (e.g. 'slightly', 'a bit') in British courtroom speech from 1700 to 1900. The study is based on the 24-million-word Old Bailey Corpus (OBC 2.0), which can be supplemented by the complete material from the Old Bailey Proceedings available online (134 million words). It exploits available corpora and aims at the consolidation of methods in historical pragmatics and historical sociolinguistics. The project is carried out in collaboration with Claudia Claridge (b. 1965) at the University of Augsburg and Ewa Jonsson (b. 1968) at Mid Sweden University and Uppsala University.

Although corpora are widely used nowadays, Kytö feels that financing bodies have tended to be relatively restrictive in granting funding to corpus builders, as corpus compilation has not been considered real research but more of an activity contributing to research infrastructure. An improvement occurred when the Council for Research Infrastructures (RFI) was created in 2001, but here too corpus builders have encountered problems when competing with applicants from the natural and life sciences.

Merja Kytö, like Jan Svartvik and Gunnel Engwall, was brought into corpus linguistics during her doctoral studies. The Helsinki Corpus project was considered radical in Finland at the time (the 1980s) and was ultimately acknowledged. However, in contrast to Sweden, corpus linguistics did

---

70 Grund's doctoral thesis was an edition of Humfrey Lock's *Treatise on Alchemy* (Grund, 2004, published as Grund, 2011), while Terry Walker dealt with second-person singular pronouns in Early Modern English dialogues (Walker, 2005, published as Walker, 2007).

not play such a significant role in other language departments in Finland. Nevertheless, the Department of English at the University of Helsinki managed to secure long-term financing through a centre of excellence-grant. Although this funding had significant effects, the case also shows the difficulties in the long run. In the course of time, excellent institutions run the risk of being considered normal, since other institutions are adopting the same type of research. In addition, funding bodies have an unwillingness to support the same type of research or the same institution for many years. Finally, the Kytö case underlines the international character of the field. She has moved to Sweden in order to pursue her research and she has taken an active part in the work of an international organization.

Conclusions

The two cases we have presented in this chapter demonstrate how conditions for corpus linguistics changed after the 1970s. Technological developments clearly facilitated and sped up the creation of corpora, and as a result the use of corpora has become much more widely accepted.

In terms of the Swedish language there is no doubt that the work that Sture Allén started more than fifty years ago has become highly institutionalized in *Språkbanken* (containing vast volumes of modern text), for which Lars Borin is now responsible. Together with *Litteraturbanken* (consisting of an increasing share of the Swedish literature) and SWE-CLARIN, it represents an invaluable source for what today is often referred to as the digital humanities. Likewise, the work that Jan Svartvik started as a doctoral student at Uppsala University is now continued by Merja Kytö.

Both Borin and Kytö began work in the area of corpus linguistics early in their careers. They were both recruited by their doctoral supervisors. In this way their examples point to the importance of the recruitment of doctoral students to emerging fields. In both cases the supervisors can be considered scientific entrepreneurs who embarked on new research journeys. Borin

and Kytö also underline the need for close cooperation with experts in the computer centres at their universities.

Obviously, the financing of the research has been a key factor. For both Borin and Kytö the Swedish Research Council (and its predecessors) and the Bank of Sweden Tercentenary Foundation have been particularly important. For Kytö, the Academy of Finland, which is the Finnish Research Council, was also crucial to her work at the University of Helsinki. However, it should also be noted that they have both felt some resistance towards the financing of new corpus production, something which has to a certain extent been remedied by the special allocations for infrastructure at the Swedish Research Council.

Finally, it is worth mentioning that both Borin and Kytö are significant actors internationally. Through *Språkbanken*, Borin is leading the Swedish node of the European Union initiative, the CLARIN network, while Kytö is secretary and co-editor within the international network of corpus linguists through ICAME (International Computer Archive of Modern and Medieval English). As will be demonstrated in Chapter 8, this is just one of the many organizations that have appeared in the field over the years. These in turn are signs of the institutionalization of the field.

CHAPTER 7. A SECOND GENERATION DEALING WITH SPOKEN LANGUAGE

Introduction

The previous chapter demonstrated that technical development has made it much easier for the second generation of corpus linguists dealing with written language. Obviously, the same changes were also important for the handling of spoken material. As pointed out in Chapter 2, such work got underway as early as the late 1950s by Randolph Quirk, in due course joined by others such as Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. The continuation of these traditions will be dealt with in the present chapter, using the cases of Åke Viberg and Jens Allwood as examples. Interestingly, both have had a background in studies of the work of Noam Chomsky.

From generativist to second-language acquisition: Åke Viberg in Stockholm

Åke Viberg (b. 1945) started as a generative linguist in the early 1970s at the Department of Linguistics at Stockholm University, teaching for a number of years at the undergraduate level.[71] During this period, together with a colleague, he published a general introduction in Swedish to Chomsky's grammar built on intuition as a method for interpretation (Trampe & Viberg, 1972). This work was used for a number of years as a textbook in Sweden.

---

71 This section is mainly based on an interview with Åke Viberg, by Tina Hedmo on August 29, 2011 with additional information from Åke Viberg in March 2018.

Around 1974, he turned to studies of Swedish as a second language, which at the time was a rather small research area, both in Sweden and internationally. The change in research profile, from what Fillmore (1992) has termed an 'armchair linguist' to a 'corpus linguist', required new resources for the collection of data. The start for Viberg was a quite extensive, not yet digitalized, corpus of essays in Swedish written by foreign students, covering ten different native languages with ten writers within each language. The leader of the project, Björn Hammarberg (b. 1935), later converted it into a digital and modern form of corpus. Viberg's main responsibility was to work out a comparative description of the ten first languages and Swedish based on published grammars and on translations of Swedish examples into these languages. This project resulted in a number of joint publications, the best-known being Hammarberg & Viberg (1977).

The research was funded for eight years by the Swedish National Agency for Education (*Skolöverstyrelsen*), and Viberg was engaged in the project as a research assistant. During his PhD-studies he spent a short period at Lund University and participated in a number of international conferences. One of these was a conference in the area of second-language acquisition arranged for researchers in Germany and the Nordic countries in Berlin in the late 1970s. At this event, he met some of the leading researchers studying German as a second language for immigrants, which was important for the development of his research.

In 1981 Viberg completed his PhD in contrastive lexicology (Viberg, 1981) based on the project on Swedish as a target language (*Svenska som målspråk*, SSM) with Östen Dahl as his final supervisor after his first supervisor Bengt Sigurd moved to a chair at Lund University (see Chapter 4, pp. 41 and 43). In 1983 the most important results of the dissertation were published in *Linguistics*, a highly respected international journal in the field (Viberg, 1983). This article, which has been important for Viberg's career, is still cited. After his dissertation, Viberg continued his research through lexical studies with continued support from his supervisor Bengt Sigurd, for instance in relation to an application to the Research Council for the Humanities and the Social Sciences.

Through his project on target languages, Viberg came in contact with colleagues in the other Nordic countries, mainly because the topic was becoming equally relevant in Denmark and Norway. Hence Swedes in a way became forerunners in this kind of research on second-language acquisition, followed first by Norwegians and later on by Danes. In Sweden, Viberg worked with researchers at the Department of Nordic Languages at the University of Gothenburg, who were running another large Swedish project studying language development among immigrants. In contrast to Viberg's own research objects (foreign students), they studied schoolchildren.

In the mid-1980s, Viberg gained a position as Associate Professor at the Centre for Research on Bilingualism at Stockholm University. There he was the principal coordinator of large projects recording and translating speech of more than a hundred individuals, both the language of schoolchildren and second-language acquisition, using adults as the population. In the case of children's speech, recordings were the most appropriate way to study such language. The material was transcribed and systematized in a chronological order. Doctoral students were involved in these projects for gathering data, and their work resulted in a number of doctoral dissertations.

During this time, Viberg started to use computers more frequently, although these were simple and slow in comparison with computers of today. However, they still facilitated analyses on a much larger scale and opened up for new forms of investigation. At this time, the texts were stored on tape. There were errors, but nevertheless it saved time.

In terms of funding of the research in the 1980s, Viberg received support from the National Swedish Board of Health and Welfare (*Socialstyrelsen*) and the Social Welfare Board of the Stockholm suburb of Rinkeby.[72] These resources were a necessity, since the recordings and transcriptions were very resource-demanding. Another source of funding that Viberg found central for this applied research was the Research Council for the Humanities and

72 Rinkeby is a suburb northwest of Stockholm, with an overrepresentation of immigrants.

the Social Sciences, although its support was more uncertain, especially as second-language acquisition was looked upon as a low-status area, occupying a more peripheral position in the field of linguistics. Today the area is well established and organized within the Centre for Research on Bilingualism, and it has gained considerable resources from both the Bank of Sweden Tercentenary Foundation and the Swedish Research Council.

In 1994, after spending ten years as an associate professor in Stockholm, Viberg applied for a chair in general linguistics at Lund University, which was attractive to him since colleagues there were dealing with bilingual research, second-language acquisition and typology. Apart from that, Lund had about ten PhD-students doing research in this area. Viberg had also already participated in research projects in Lund, one being a continuation of his earlier work in comparing languages. He was therefore well connected to the Lund environment when he was appointed to the chair in 1994. For a period, he was also Head of Department. However, in 2001 he moved to Uppsala where a chair was available in general linguistics.

In later years, Viberg has concentrated on building small and specific corpora for use in studying, among other things, how the Swedish language is organized and structured semantically. He has also studied the semantics of Swedish verbs in a comparative perspective, with an emphasis of neighbouring languages, thereby collaborating informally with linguists and experts in these languages. An important reason for not building large, new corpora was the fact that several large multilingual corpora were becoming available on the Internet. In addition, he already had his own material to work on (his learner corpora, the translation corpus and other small and topic-specific corpora).

Åke Viberg is an interesting example of a scholar who has tried new approaches. Starting out in the tradition of Chomsky, he has moved towards corpus linguistics, although he has not completely giving up what he learnt from Chomsky. He has also entered a new field of linguistics by focusing on second-language acquisition, for which he initially met resistance but in the end found strong acceptance. A third special feature of his case is the

financing of the research from both a state agency not primarily financing research and the community where he undertook his studies. Once again we can see how non-traditional funding has contributed to innovative research.

From philosophy to analysis of spoken language and multimodal communication: Jens Allwood in Gothenburg

Jens Allwood (b. 1947) has an academic background in linguistics in an environment hostile to statistics and frequencies in the 1960s and 1970s.[73] He moved into corpus linguistics in the 1980s and has since built and used corpora in varying research areas like spoken language, communication and empirical concept analysis. The movement to corpus linguistics forced him to learn and develop new quantitative techniques, methods and, to some degree, new theories.

Apart from doing corpus-related research, Jens Allwood has dealt with semantics, pragmatics, intercultural and interdisciplinary communication. As such, his research crosses various sub-fields and areas in linguistics and communication studies, and his scope has extended over time. In addition, most of his projects have been run simultaneously.

In the 1960s, before he started his doctoral education, Allwood studied a number of subjects in parallel with linguistics at the University of Gothenburg, such as sociology and philosophy. One of his main sources of inspiration was the intellectual and open-minded professor of theoretical philosophy, Ivar Segelberg (1914–1987), who raised questions in a provocative way. Allwood also came in contact with Per Lindström (1936–2009), one of Sweden's most internationally renowned mathematical logicians.

---

73 This section is mainly based on an interview with Jens Allwood, by Tina Hedmo on October 1, 2011 with additional information from Jens Allwood in March 2018.

When Allwood asked Lindström whether it was not possible to study logic outside of mathematics, in order to find out how people reason logically in different cultures, Lindström told Allwood to look at Noam Chomsky, who, according to Lindström, believed all human languages had a common deep structure. Allwood then read Chomsky's *Syntactic Structures* (Chomsky, 1957), *Aspects of the Theory of Syntax* (Chomsky, 1965) and *Cartesian Linguistics* (Chomsky, 1966). These readings inspired Allwood to write a C-level essay in theoretical linguistics, in which he evaluated the philosophical interest of Chomsky's ideas (Allwood, 1969), which he claimed mainly amounted to a reawakening of the rationalist doctrine of innate ideas and possibly a type of linguistic neo-Kantianism.

After spending a year at the University of Massachusetts at Amherst in 1974, two years later he defended his thesis *Linguistic Communication in Action and Cooperation* (Allwood, 1976) in the Department of Linguistics at the University of Gothenburg. He then applied for a post at Uppsala University, where he knew, from earlier contacts at a summer school in linguistics, that the professor, Sven Öhman, shared his scepticism of Chomsky (Öhman, 2007). After getting the post in Uppsala, Allwood spent five years (1976–1980) as a senior lecturer and director of studies in the Department of Linguistics there. His main research at the time was devoted to semantics and pragmatics albeit with a philosophic touch, even though phonetics was the main area of research at Uppsala at this time.

In 1978 Allwood, together with the Gothenburg social anthropology professor Göran Aijmer (b. 1936), started a research project called *Anthropological Linguistics*. It was a valuable project for Allwood, funded by the Research Council for the Humanities and the Social Sciences. After one year each at Linköping University and Stockholm University as a *docent*, he returned to Gothenburg in 1982 as an associate professor and later also head of department. This gave him an opportunity to arrange a summer school in Artificial Intelligence, which opened up for discussions regarding the combination of computers and language. The school also raised the idea of starting an undergraduate programme in language and computers, later named computational linguistics, which took about two years to prepare.

In this process Allwood and his colleagues involved Sture Allén at *Språkbanken*, and the programme started in 1984. In the meantime, a chair in linguistics had been created and Allwood became its first holder in 1986. At the turn of the 1980s, he continued his collaboration with other departments by starting the interdisciplinary centre *Språk, Semantik, Kognition, Kommunikation, Interaktion och Information* (SSKKII).[74]

In the 1980s, Allwood also became interested in multimodality, which led to a number of projects. In this research, conversations of individuals were video recorded, whereupon data was transcribed and stored as large spoken computerized corpora. Allwood's interest in this research area had its roots in the belief that linguistics was far too strongly focused on written language. In addition, he firmly believed that the basis of human communication was face-to-face communication, which demanded video recording.

In the late 1980s, Allwood coordinated a large corpus project in spoken language learning, *Ecology of Adult Language Acquisition*. Among other things, this project led to research contacts with Åke Viberg in Uppsala and Björn Hammarberg in Stockholm (cf. above p. 72).

In processing the data, Allwood and his colleagues found that virtually no appropriate software programs existed. The work therefore also involved a lot of learning, program development and standardization. For instance, it was apparent that people transcribed differently, entailing that the researchers needed to find a common framework for how to transcribe recorded data.

Another idea raised by Allwood at this time was that language varies with social context. He believed in the necessity of studying spoken language in real-life situations rather than in studios. Consequently, Allwood and his colleagues recorded speech from about thirty different social activities. Here, too, the researchers used the corpus that resulted from these recordings for various purposes, and they had to construct software programs

74 The aim of SSKKII is to organize both theoretically and practically oriented research projects. SSKKII provides a link between research projects, industry and trade.

for data processing. These projects in turn led to research contacts all over the world, especially as computational linguistics was a new sub-discipline in linguistics.

In later years, Allwood has returned to the area of multimodal communication, particularly the use of gestures, and is involved in the compilation of a multimodal corpus, concentrating on sight and hearing. This corpus will be added to the old spoken language corpora, as these areas are closely related. The results of the research have been published in a number of articles over the last few years.[75]

All through his career Allwood's research has received grants from external funding bodies. This has been particularly important, since his research in spoken language is resource-demanding in terms of equipment, technical know-how and disciplinary knowledge, that is, recording, transcribing and coding the empirical material. In addition to computers, there is a need for cameras, microphones and good recording equipment and, of course, competent people. However, the opportunities to attract funding for his research have at times been limited. As a consequence, Allwood has sometimes played down his intention to build spoken language corpora in research applications. In addition, most of his research has been funded by smaller amounts from a variety of external funding sources.

Over the years, Allwood has had exchanges with a large number of researchers, especially outside Sweden. Early in his career he was inspired by the ideas of Emanuel Schegloff (b. 1937), an American researcher in the area of conversation analysis, who was a guest in Uppsala in the late 1970s and also a lecturer in another summer school that Allwood arranged in 1984.[76] Later on, Allwood had exchanges with Scandinavian researchers through the network NordTalk, a Nordic research network in corpus-based research on spoken language. His assignment as consulting editor of the *Journal of*

75 See e.g. Allwood (2008a and 2008b; 2013).

76 See e.g. Ochs, Schegloff & Thompson (1996); Schegloff (2006).

Corpus Linguistics has also facilitated contacts with British scholars. In Germany he has contacts in the area of spoken language with for instance the Hamburg professor Jochen Rehbein (b. 1939).

Later research, aiming at compiling spoken language corpora in China, South Africa, Malaysia and Nepal, has resulted in more contacts with corpus linguists in these parts of the world. The Nepal project aims to construct a multimodal lexicon and is run together with the largest university in Nepal and a former student at the Department of Information Technology at the University of Gothenburg. The Swedish International Development Cooperation Agency (SIDA) and the Swedish Research Council cover meeting costs for this project, but unfortunately not equipment and salaries.

Jens Allwood is a linguist who has moved between institutions and research interests. Starting out in philosophy, especially studying Wittgenstein and Kant, and later in studies of Chomsky, he turned to communication theory and corpus linguistics through studies of spoken language. In relation to the first generation of corpus linguists, it is particularly worth noting that he has followed the track blazed by Randolph Quirk and his collaborators. In addition, he linked back to the first generation through his collaboration with Sture Allén in the computational linguistics educational programme. Allwood's work has added further competence to digital humanities in Sweden in general and in Gothenburg in particular.

Conclusions

This chapter has demonstrated how two Swedish linguists, Åke Viberg and Jens Allwood, who originally were following the ideas of Noam Chomsky moved on to corpus linguistics. In the case of Viberg, it is a move into second-language acquisition research, first with written material and later on spoken language, while Allwood has developed an orientation towards interaction in communication through multimodal studies. In this way

they both related to the early researchers, such as Quirk and Svartvik, who both took an interest in spoken language. However, it is also tempting to interpret Viberg's and Allwood's interest in speech as a link to their early contacts with Chomsky's thinking.

There is no doubt that both Viberg and Allwood have had successful academic careers, although they chose non-traditional approaches. With time their research orientations have successively been taken up by others. For this to happen it is again clear that the financing of their research has been a fundamental condition. Both have received grants from the Swedish Research Council as well as from other sources, even from agencies that were not primarily research funding bodies. In Viberg's case his dissertation work was financed by the Swedish National Agency for Education (*Skolöverstyrelsen*) and his later work by the National Swedish Board of Health and Welfare (*Socialstyrelsen*) as well as the Social Welfare Board of the Stockholm suburb of Rinkeby, while Allwood had grants from the Swedish International Development Cooperation Agency (SIDA). These examples show how practical interests may help researchers to fund new ideas. It is also evident that funding from their universities as well as smaller grants from different sources have played a significant role for the accomplishment of their research.

CHAPTER 8. LATER INTERNATIONAL DEVELOPMENT

Introduction

As we have now presented eight actors on the Swedish scene, it is appropriate to return to the SciVerse Scopus search in Chapter 2, where we presented the most cited works before 1970. In this chapter will provide the results from an analysis of later international developments concerning the titles and authors in the database we developed through the search in SciVerse Scopus.[77] The results are presented as follows: the most frequent titles in the period 1970–1999, the most frequently cited authors as well as patterns of relationships over time. In addition, we will provide another indicator of the development of corpus linguistics, namely the organizing of the field through the foundation of a number of international organizations.

The most frequent titles 1970–1999

It is obvious that the decades after 1970 brought new titles fitting into the profile of our search (Table 8.1, upper part). At the top among works published in the 1970s we find *The American Heritage Word Frequency* by John Carroll, Peter Davies and Barry Richman (1971).

Second is the British semiotics linguists Michael A. K. Halliday (a student of Rupert Firth, a top reference above in Table 2.1) and Ruqaiya Hasan with *Cohesion in English* (Halliday & Hasan, 1976). Four authors of a co-authored

77 For the search criteria, see Chapter 2, p. 24.

Table 8.1. The most cited works from 1970–1979, 1980–1989 and 1990–1999 in a SciVerse Scopus search for corpus-related works

1970–1979

Carroll, John B., Peter Davies & Barry Richman, 1971, The American Heritage Word Frequency Book. New York: Houghton Mifflin.
Halliday, Michael A. K. & Ruqaiya Hasan, 1976, Cohesion in English. London: Longman.
Coltheart, Max, Eddy Davelaar, Jon Torfi Jonasson & Derek Besner, 1977, 'Access to the Internal Lexicon', in: Dornic, S. (ed.), Attention and Performance, VI. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 535–555.
Taft, Marcus, 1979, 'Recognition of Affixed Words and the Word Frequency Effect', Memory and Cognition, 7, pp. 263–272.
Forster, Kenneth I., 1976, 'Accessing the Mental Lexicon', in: Wales, Roger J. & Edward Walker (eds), New Approaches to Language Mechanisms. Amsterdam: North-Holland, pp. 257–287.
Labov, William, 1972, Sociolinguistic Patterns. Philadelphia: University of Pennsylvania Press.

1980–1989

Quirk, Randolph , Sidney Greenbaum, Geoffrey Leech & Jan Svartvik, 1985, A Comprehensive Grammar of the English Language. London: Longman.
Levelt, Willem J. M., 1989, Speaking: From Intention to Articulation. Cambridge, MA: MIT Press.
Biber, Douglas, 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.
McClelland, James L., David E. Rumelhart & the PDP research group, 1986, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume II. Cambridge, MA: MIT Press.
Seidenberg, Mark. S. & James L. McClelland, 1989, 'A Distributed, Developmental Model of Word Recognition and Naming', Psychological Review, 96, pp. 523–568.
Dell, Gary S., 1986, 'A Spreading-activation Theory of Retrieval in Sentence Production', Psychological Review, 93 (3), pp. 283–321.

1990–1999

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan, 1999, The Longman Grammar of Spoken and Written English. London: Longman.
Sinclair, John McHardy, 1991, Corpus, Concordance, Collocation. Oxford: Oxford University Press.
Levelt, Willem J. M. &Linda Wheeldon, 1994, 'Do Speakers have Access to a Mental Syllabary?', Cognition, 50, pp. 239–269.
Levelt, Willem J. M., Ardi Roelofs & Antje S. Meyer, 1999, 'A Theory of Lexical Access in Speech Production', Behavioral and Brain Sciences, 22, pp. 1–75.
Biber, Douglas, Susan Conrad & Randi Reppen, 1998, Corpus Linguistics: Investigating Language. Structure and Use. Cambridge: Cambridge University Press.

Source: See Chapter 2, p. 24.

paper (Coltheart et al., 1977) on access to the internal lexicon follow. After them we find Marcus Taft with a paper on the word-frequency effect on word recognition (Taft, 1979). Further down is the psychologist Kenneth I. Forster with a paper on the access to the mental lexicon (Forster, 1976) and the sociolinguist William Labov with *Sociolinguistic Patterns* (Labov, 1972). The 1970s thus exhibits a word-frequency dictionary at the top, which is followed by other types of linguists who make use of corpora, namely, those working with semiotics, psycholinguistics and sociolinguistics.

In the 1980s (Table 8.1, middle part) we find at the top Randolph Quirk and his collaborators with their grammar of the English language (Quirk et al., 1985). After them follow *Speaking: From Intention to Articulation* (Levelt, 1989) by the Dutch psycholinguist Willem Levelt (b. 1938) and Douglas Biber's *Variation across Speech and Writing* (Biber, 1988). These are definitely corpus users, and the same can be said about James McClelland (b. 1948) and his research group, which is represented by *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (McClelland, Rumelhart & the PDP research group, 1986) and a *Psychological Review* paper on word recognition and naming (Seidenberg & McClelland, 1989). Finally from the 1980s, there is another *Psychological Review* paper by Gary Dell on sentence production (Dell, 1986). Thus, the first title on the list in the 1980s is the result of corpus building, while the others are characterized by corpus use. It is particularly worth noting that two papers from a psychology journal have made it to the top.

In the 1990s (Table 8.1, bottom part) the top title is *The Longman Grammar of Spoken and Written English* (Biber et al., 1999), i.e. a title with a strong link to corpus building. This is also the case for the British lexicographer John Sinclair and his *Corpus, Concordance, Collocation* (Sinclair, 1991). These are followed by two entries of the above-mentioned Dutch psycholinguist Willem Levelt (Levelt & Wheeldon, 1994; Levelt, Roelofs & Meyer, 1999). Last on the list is a co-authored book on corpus linguistics (Biber, Conrad & Reppen, 1998). Thus, In the 1990s works of corpus builders were more successful in being frequently cited than before.

The above implies that we can see among the top references a number of corpus builders, who presented their corpora and analyses thereof. It has also been evident that the corpus building has been important to other researchers, who have been able to use corpora in their work: psycholinguists, sociolinguists, as well as scholars working with semiotics and phonetics.

The most-cited authors

It has been evident in the previous section that some actors appear with more than one work at the top. We could also expect that they have a number of publications which are outside the top lists. It is therefore appropriate to also look at the most-cited authors in the material. In so doing, we found 23 authors in 2010, each with more than 200 citations (Table 8.2). These authors can be divided into two groups, thirteen of whom are linguists and ten are psychologists or scholars of cognitive science.

The linguists are Geoffrey Leech (506), Douglas Biber (430), Susan Conrad (319), John Sinclair (307), Harald R. Baayen (275), Michael A. K. Halliday (273), Stig Johansson (269), Joan Bybee (263), Nelson W. Francis (260), Edward Finegan (237), Sidney Greenbaum (220), Jan Svartvik (216), and Randolph Quirk (215). Clearly, men dominate the group: only two of the top cited linguists are women (Susan Conrad and Joan Bybee). There is also a majority (eight out of thirteen) affiliated with European institutions. In terms of their age, the median year of birth is 1937, with 1911 (Nelson W. Francis) being the earliest and 1960 (Susan Conrad) the latest year of birth.

In the group of psychologist and scholars of cognitive science we find Mark Seidenberg (344), James McClelland (321), Willem Levelt (290), Michael A. K. Halliday (273), Kenneth Forster (263), William Marslen-Wilson (263), Alfonso Caramazza (251), Max Coltheart (250), Kevin Patterson (238), Brian MacWhinney (221) and David Balota (211). All of them are men, and six of the ten are associated with universities in the United States. On the whole they are younger than the members of the linguist group: the median

Table 8.2. Authors with more than 200 citations in a SciVerse Scopus search for corpus-related works in 2010

Author	Institution	Country	Expertise	Citations
Leech, Geoffrey (1936–2014)	Lancaster University	UK	Linguistics and modern English	506
Biber, Douglas (b. 1952)	Northern Arizona University	USA	Applied linguistics	430
Seidenberg, Mark (b. 1953)	University of Wisconsin-Madison	USA	Psychology and cognitive neuroscience	344
McClelland, James (b. 1948)	Center for Mind, Brain and Computation, Stanford University	USA	Psychology	321
Conrad, Susan (b. 1960)	Portland State University	USA	Applied linguistics	319
Sinclair, John (1933–2007)	Birmingham University	UK	Modern English language	307
Levelt, Willem (b. 1938)	Max Planch Institute for Psycholinguistics	NL	Psycholinguistics	290
Baayen, R. Harald (b. 1958)	University of Tübingen and University of Alberta	D	Quantitative linguistics	275
Halliday, Michael A. K. (b. 1925)	University College London and University of Sydney	UK, AU	Linguistics	273
Johansson, Stig (1939–2010)	University of Oslo	N	Modern English	269
Forster, Kenneth I. (1945)	University of Arizona	USA	Psychology	263
Bybee, Joan L. (b. 1945)	University of New Mexico	USA	Morphology, phonology	263
Marslen-Wilson, William (1945)	MRC Cognition and Brain Sciences, University of Cambridge	UK	Cognitive science, neuroscience	263
Francis, Nelson W. (1911–2002)	Brown University	USA	Corpus linguistics	260
Caramazza, Alfonso (1946)	Harvard University	USA	Psychology	251
Coltheart, Max (b. 1939)	Macquairie University	AU	Cognitive science	250
Patterson, Kevin (1968)	University of Leicester	UK	Psychology	238
Finegan, Edward (b. 1940)	University of Southern California	USA	Linguistics and Law	237
MacWhinney, Brian (b. 1945)	Carnegie-Mellon University	USA	Psychology	221
Greenbaum, Sidney (1929–1996)	University College London	UK	English language and linguistics	220
Svartvik, Jan (b. 1931)	Lund University	SE	English	216
Quirk, Randolph (1920–2017)	University College London	UK	English	215
Balota, David (b. 1954)	Washington University	USA	Psychology and neurology	211

Source: See Chapter 2, p. 24.

year of birth is 1946, with 1938 and 1968 being the extreme values (Willem Levelt and Kevin Patterson, respectively).

The two clusters we have obtained in our search in 2010 can be further illustrated by a co-citation chart between authors with co-citations numbering 20 or more, produced by means of the Pajek data program (Figure 8.1). It shows nicely how the linguists are linked together in the right-hand cluster, where we find (in alphabetical order): Biber, Conrad, Finegan, Greenbaum, Halliday, Johansson, Leech, Quirk, Sinclair and Svartvik. Then as a link between the two clusters is Joan Bybee, who does research on morphology and phonology. Interestingly enough, there is one strong corpus linguist in the left-hand psychology-oriented cluster (Nelson W. Francis) as well as a quantitative linguist (Harald R. Baayen). However, otherwise the members of the left-hand cluster are psychologists or cognitive scientists, again in alphabetical order: Balota, Caramazza, Coltheart, Forster, Levelt, Marslen-Wilson, MacWhinney, McClelland, Patterson and Seidenberg.

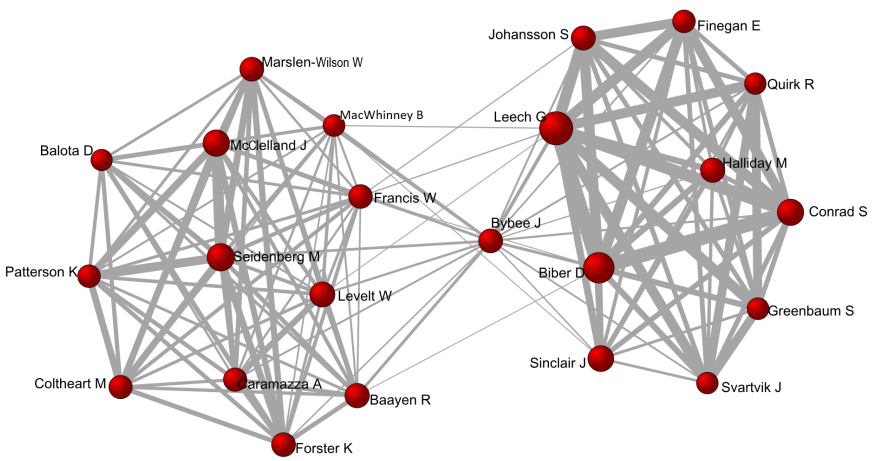


Figure 8.1. Co-citations among the top authors in 2010.

Table 8.2 and Figure 8.1 show that our broad search until 2010 using a number of search items has provided a database of references and authors associated with both corpus building and corpus use. A more limited search just using 'corpus linguistics' thus excludes the left-hand cluster in Figure 8.1, that is, the corpus users. In such a search the authors with citations above 200 are: Biber (627), Leech (615), Conrad (478), Sinclair (407), Halliday (362), Johansson (322), Finegan (294), Greenbaum (262), Quirk (254), Svartvik (253), Reppen (246), McEnery (228), Stubbs (216), Hunston (204), and Wilson (203). Of these, the first ten (from Biber to Svartvik) are included in our earlier top list, but now with higher citation counts. Of the remaining five, Reppen worked with Biber and Conrad (cf. e.g. Biber, Conrad & Reppen, 1998), McEnery has published an introduction to corpus linguistics (McEnery & Wilson, 1996) and another text on corpus-based language studies (McEnery, Xiao & Tono, 2006), while Stubbs has published a book on corpus studies of lexical semantics (Stubbs, 2002) and Hunston & Francis (2000) deals with a corpus-based lexical grammar. Wilson, finally, is McEnery's co-author. Hence, the more restricted search confirms the earlier top positions but also provides some additional frequently cited scholars in corpus linguistics.

Development over time

Our SciVerse Scopus search has also made it possible to map the development of the field through an analysis in which the five most cited authors were selected for each decade. An additional requirement was that the number of years between the pairs in a co-citation should be longer than ten years. The result of this analysis is the mapping exhibited in Figure 8.2.

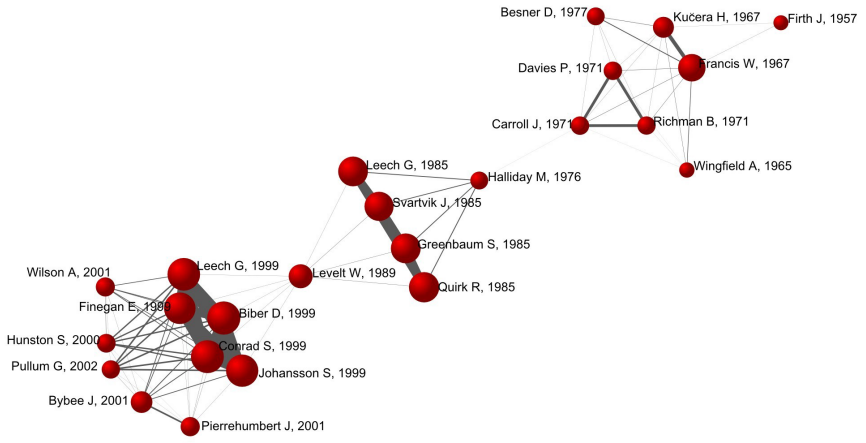


Figure 8.2. Development of the field until 2010.

In the first cluster (upper right in Figure 8.2) we can see the creators of the Brown corpus (Kučera & Francis, 1967) as central. They build on the work of John Rupert Firth (1957, who was a forerunner in linguistic research (cf. Chapter 2, p. 25). We also see in the first cluster the strong triangle of Carroll, Davies and Richman and their *The American Heritage Word Frequency Book* (1971). In addition, the first cluster includes Arthur Wingfield, co-author of the Oldfield & Wingfield (1965) paper as well as Derek Besner, one of the authors of the Coltheart et al. (1977) paper. These two indicate the early links to psychological research.

The mid-cluster (Figure 8.2, middle) is dominated by Randolph Quirk and his collaborators with their English grammar *A Comprehensive Grammar of the English Language* (Quirk et al., 1985). They have a link back to the classical linguist Michael A. K. Halliday and forward to the psycholinguist Willem Levelt and his 1989 book *Speaking: From Intention to Articulation*. He is also the link to the last cluster (Figure 8.2, lower part) where the authors of another English grammar *The Longman Grammar of Spoken and Written English* (Biber et al., 1999) are central. They are linked to Geoffrey K. Pullum, co-author of another grammar, *The Cambridge Grammar of the*

English Language (Huddleston & Pullum, 2002), Joan Bybee, co-editor of *Frequency and the Emergence of Linguistic Structure* (Bybee & Hopper, 2001), to Janet Pierrehumbert a contributor to the same volume (Pierrehumbert, 2001), Susan Hunston, co-author of *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* (Hunston & Francis, 2000), and finally Andrew Wilson, co-author of *Word Frequencies in Written and Spoken English: Based on the British National Corpus* (Leech, Rayson & Wilson, 2001). The last cluster thus gives evidence of a strong representation of grammars and corpora. The representation of the psychologists is meagre, pointing to the fact that the selection criteria have implied that we have been focusing on the right-hand cluster of Figure 8.1. Finally, it should be mentioned that the development from the right-hand cluster to the left-hand cluster means higher citations (larger red circles) and closer relationships (a tighter network).

The organizing of the field

Our analysis of the SciVerse Scopus data above has shown how the corpus builders and the corpus users are linked together. Another indicator of the linkages within the field is the international organizing, which we have studied within the project and presented in a separate paper (Engwall & Hedmo, 2016). In so doing, we have even been able to formulate a more general model for the organizing of scientific fields (Figure 8.3).

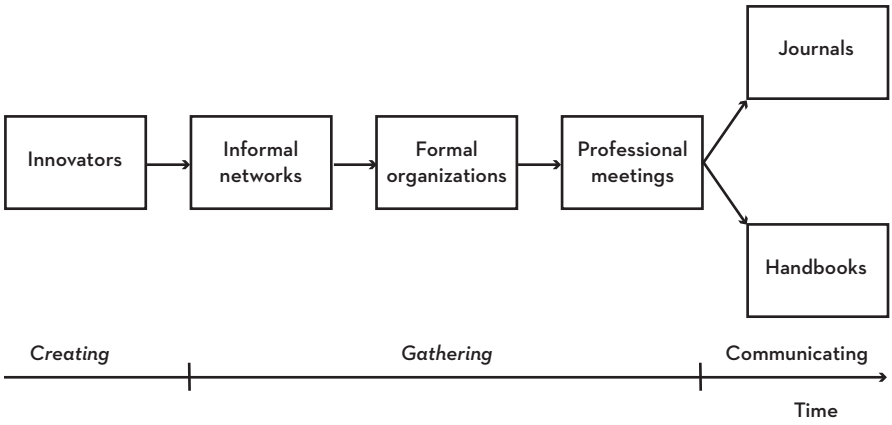


Figure 8.3. A model for the organizing of scientific fields.

Our model is based on the observation that scientific innovations emerge in many different countries at the same time and that scientific entrepreneurs eventually find colleagues in other countries who think along the same lines. Over time, in a creation phase, they will build informal networks with seminars, workshops, summer schools, etc. After some time, this in turn will lead to a gathering phase, during which formalized organizations are created and professional meetings are arranged. Then, the new field shifts into a communicating phase, which includes the launching of journals and other means of communication. An ultimate sign of the establishment of a field is then the publication of handbooks.

In terms of the gathering phase we found (Table 8.3) that, as mentioned in Chapter 2, for corpus builders that the development started in the 1960s through the foundation of two organizations in the United States: the Association of Computational Linguistics (ACL) and the International Committee on Computational Linguistics (ICCL).

ACL and ICCL were in the 1970s followed by three European organizations and one North American: the International Computer Archive of Modern and Medieval English (ICAME) in 1969, the Association for Literary and Linguistic Computing (ALLC) in 1973, the Nordic Association of Linguists (NAL) in 1976 and the Association for Computing and the

Table 8.3. Organizations in the field of corpus linguistics

Organization	Acronym	Founded	Area
Association for Computational Linguistics	ACL (AMTCL)	1962	North America
International Committee on Computational Linguistics	ICCL	~1965	Transnational
International Computer Archive of Modern and Medieval English	ICAME	1969	Europe
Association for Literary and Linguistic Computing	ALLC	1973	Europe
Nordic Association of Linguists	NAL	1976	Europe
Association for Computing and the Humanities	ACH	1978	North America
European Association for Lexicography	EURALEX	1983	Europe
Canadian Society for Digital Humanities	SDH-SEMI	1986	North America
International Association for Machine Translation	IAMT	1991	Transnational
Linguistic Data Consortium	LDC	1992	North America
European Corpus Initiative Multilingual Corpus	ECI/MCI	1992	Europe
Alliance of Digital Humanities Organizations	ADHO	2005	Transnational

Source: Engwall & Hedmo (2016), Tables 1–3.

Humanities (ACH) in 1978. Another two appeared in the 1980s: the European Association for Lexicography (EURALEX) in 1983 and the Canadian Society for Digital Humanities (SDH-SEMI) in 1986, while the 1990s saw the appearance of the transnational International Association for Machine Translation (IAMT) in 1991, the North American Linguistic Data Consortium (LDC) in 1992 and the European Corpus Initiative Multilingual Corpus (ECI/MCI), also in 1992. Another transnational association, the Alliance of Digital Humanities Organizations (ADHO) was founded in 2005. In addition to these organizations there have obviously been other initiatives to gather scholars in the field of corpus linguistics in a narrow as well as in the wider sense of cognitive science.

Likewise, we found that the majority of the organizations have started journals, with *Computational Linguistics* launched in 1974 by the ACL being the first (Table 8.4). It was followed in 1978 by journals launched by ICAME and NAL: the *ICAME Journal* and the *Nordic Journal of Linguistics*. Similarly, in the 1980s, ALLC, ACH and SDH-SEMI jointly started *Literary & Linguistic Computing* in 1986, and EURALEX launched the *International Journal of Lexicography* in 1988, both with Oxford University Press as publisher. In the 1990s and 2000s commercial publishers had realized the potential of the field: in 1995 John Benjamins thus started the *International Journal of Corpus Linguistics*, followed in 2005 and 2006 by Springer, de Gruyter and Edinburgh University Press publishing *Language Resources and Evaluation*, *Corpus Linguistics and Linguistic Theory*, and *Corpora*, respectively. In 2007 the transnational organization ADHO started the publication of the *Digital Humanities Quarterly*.

Table 8.4. Significant journals of the field

Journal Name	Start	Organization	Publisher
Computational Linguistics	1974	ACL (AMTCL)	MIT Press
ICAME Journal	1978	ICAME	1978–2014: Lancaster University; 2014–: de Gruyter
Nordic Journal of Linguistics	1978	NAL	Cambridge UP
Literary & Linguistic Computing	1986	ALLC, ACH, SDH-SEMI	Oxford UP
International Journal of Lexicography	1988	EURALEX	Oxford UP
International Journal of Corpus Linguistics	1995	–	John Benjamins
Language Resources and Evaluation (<2005: Computers and the Humanities)	2005		Springer
Corpus Linguistics and Linguistic Theory	2005	–	de Gruyter
Corpora	2006	–	Edinburgh UP
Digital Humanities Quarterly	2007	ADHO	ADHO

Source: Engwall & Hedmo (2016), Table 4.

Further evidence of the organizing of the field, in accordance with our model in Figure 8.1, is two handbooks (Lüdeling & Kytö, 2008 and 2009; O'Keeffe & McCarthy, 2010) published since the turn of the century. There are also today a large number of books introducing corpus linguistics (e.g. Tognini-Bonelli, 2001; Meyer, 2002; Halliday, 2004; Williams, 2005; Teubert & Čermáková, 2007; McEnery & Hardie, 2011). The latter is an additional sign that the field has made its way into an established academic discipline.

Conclusions

The analysis in this chapter has identified a number of significant works and authors, most of them men, cited in the period 1970–1999. We have seen that there is a strong cluster of researchers dealing with corpora which they have used for the development of English grammars. We have also identified another group of researchers, who can be labelled corpus users rather than corpus builders. They are cognitive scientists often with an interest in language acquisition and language loss, for whom frequency data are highly important.

In relation to Swedish developments, it is of course particularly worth noting that only one of the actors in the previous four chapters has appeared in this analysis: Jan Svartvik. His appearance in the data is also highly expected since he worked closely with Randolph Quirk and corpus linguists around him. However, the others do not appear, despite the fact that our search algorithm contained non-English words such as the French 'statistique lexicale', 'vocabulaire', 'dictionnaire des fréquences', the German 'Frequenzwörterbuch', and 'Häufigkeitswörterbuch', and the Swedish 'ordfrekvenser' and 'frekvensordbok'. This may be seen as a sign of larger interest in English as a language that has become the modern *lingua franca*, academically as well as commercially. Such an interpretation is supported by a closer look at the nationalities of the actors we have identified in this chapter, which reveals a dominance of Anglo-American researchers. The only exceptions are Harald R. Baayen of the University of Tübingen,

Willem Levelt at the Max Planck Institute in Nijmegen, Stig Johansson of Oslo University and Jan Svartvik of Lund University. Needless to say, the Anglo-American dominance may also be the result of a bias in the underlying data base towards publications in English.

The above implies that it is appropriate to go beyond bibliometric search in studies of scientific fields, particularly in the humanities and the social sciences. As demonstrated above, there has been considerable work going on in corpus linguistics for the German, French and Swedish languages. The Swedish research by Sture Allén and his successors is particularly worth mentioning here. The work he started in the 1960s is now a research programme which has been going for more than fifty years and which has resulted in a large number of databases and publications, among them word lists and grammars for Swedish. In this way it has had a significant impact on research on the Swedish language. As mentioned, the reason for the absence of this research in the international databases is likely to be found in the fact that research on a minority language in the world has difficulty penetrating the world scene. However, it may be noted that the Swedes share this situation with the French and the Germans, who also have developed very large databases of their own languages for a very long time. And, of course, studies of non-English languages are just as important as the study of English. It can even be argued that such research is even more important, since the commercial interest in them can be expected to be lower. Therefore, the need for support for research on non-English languages should be considered urgent.

CHAPTER 9. CONCLUSIONS

Conditions for scientific innovation

The development of corpus linguistics can be seen as an effort to carry out more systematic studies of various languages. While traditional linguists to a large extent excerpted texts in order to find examples for dictionaries and grammars, corpus linguists gather a large number of texts for their analysis. The purpose of this volume has been to present the results of a study of this change in language studies. For this analysis it has been found particularly appropriate to point out the significance of (1) institutional conditions and of (2) disciplinary conditions. Among institutional conditions, we expected strong authority structures to be negative for innovation, while opportunities for external funding were expected to work in the other direction. Similarly, among disciplinary conditions, nationally, strongly established approaches were expected to hamper innovation and international developments to open up for new ideas. As this model has been applied to the context of Swedish linguists we have seen that:

The international development of corpus linguistic has been going on since the last part of the nineteenth century and the first decades of the twentieth century. However, the creation of corpora took off internationally in the 1960s through the development of computer technology that facilitated the processing of large databases. Therefore, it seems fair to say that to a significant degree modern corpus linguistics is an innovation based on the availability of new technology (Chapter 2).

Authority structures in earlier days implied a concentration in a small number of universities in Sweden with professors who had considerable power within their departments. With time, authority structures have changed through the addition of several new institutions as well as an expansion of departments, which has entailed that individual professors exercised less power over their departments (Chapter 3, pp. 31–33).

External funding has a long tradition in Sweden, with research councils

that were created from the 1940s onwards as well as some important private foundations. For the field of corpus linguistics, the creation of the Bank of Sweden Tercentenary Foundation has been particularly important. However, the Swedish Research Council and its predecessor as well as other state agencies have also been supportive (Chapter 3, p. 33–36).

Established approaches in Swedish language research largely implied a focus on historical linguistics and philology. Phonetics made its way into Swedish universities through chairs in the 1950s, followed by professorships in general linguistics in the 1960s. In this way the ideas of Noam Chomsky penetrated Swedish language research (Chapter 4).

A first generation of Swedish corpus linguists

Against this background, our research has identified two generations of innovators in Sweden for corpus linguistics. The first generation (Chapter 5) includes:

Sture Allén (b. 1928), who, after completing his thesis, a commentated edition of seventeenth-century letters, in the Department of Nordic Languages at the University of Gothenburg in 1965, started a research group to study modern Swedish by means of computers. This project, which was financially supported by both the Bank of Sweden Tercentenary Foundation and the Council for Research in the Humanities, developed into a large and long-lasting research programme. In 1972 Allén was given a chair in computational linguistics, which he held until his retirement in 1993. However, his former department is still a significant node in the corpus linguistics of Swedish through the Language Bank (*Språkbanken*), the Literature Bank (*Litteraturbanken*) and SWE-CLARIN. The first institution has been particularly important for the development of Swedish dictionaries and grammars.

Jan Svartvik (b. 1931) became a pioneer in corpus linguistics through his

collaboration with Randolph Quirk at University College London within the project Survey of English Usage, work that provided the basis for his doctoral dissertation at Uppsala University in 1966. After this he continued to work with Randolph Quirk, Sidney Greenbaum and Geoffrey Leech, a collaboration that led to a number of publications, among them a frequently cited grammar. Another significant part of the collaboration was the London-Lund Corpus of Spoken Language. Again, the Bank of Sweden Tercentenary Foundation funded the research.

Inger Rosengren (b. 1934) turned to corpus linguistics after a dissertation in 1966 on adjectives in Middle High German. Inspired by her external examiner at the thesis defence, Sture Allén, she took advantage of the changes in newspaper technology, namely, the possibility of accessing typesetting tapes. For her part the corpus contained material from *Die Welt* and *Süddeutsche Zeitung*. Her research was financed by the Council for Research in the Humanities.

Gunnel Engwall (b. 1942), like Svartvik, was brought into corpus studies during her doctoral studies. Financed by her department through an assistantship and later on a doctoral scholarship, she developed a corpus of half a million words from 25 modern novels. In addition to the dissertation, this project led to a frequency dictionary and a number of published papers. This corpus was later on followed by a number of other corpora in the Department of Romance Languages at Stockholm University. She is now the chair of the above-mentioned *Litteraturbanken*.

In relation to the first generation it is worth noting that they do not appear to have met much resistance from the established professors, that is, those in power within authority structures and representing established approaches. The critique seems more to have come from the relatively new departments of linguistics. Their negative attitude does not appear to have influenced the funding decisions, however. All four were quite successful in gaining financial support for their research.

A second generation of Swedish corpus linguists

In terms of the second generation of innovators we have identified two groups, those dealing with written language (Chapter 6) and those dealing with spoken language (Chapter 7).

Our first case in the first group is *Lars Borin* (b. 1957). He was initially a student in a Slavic languages department, but later transferred to a linguistics department. After a dissertation on morphological regularities in 1991, he was involved in projects on computer-supported language teaching and learning as well as machine translation and interpretation in the 1990s. The latter introduced him to work with corpora, and this was even more the case as he moved to Gothenburg, where he became head of *Språkbanken*, once created by Sture Allén, and later on active in *Litteraturbanken* and SWE-CLARIN. In this way he is now heavily involved in corpus linguistics.

Our second case in Chapter 6 was *Merja Kytö* (b. 1953), who came into corpus linguistics during her doctoral studies at the University of Helsinki, where she took part in the compilation of a corpus of Old English texts. Although this project was well regarded in the community, and even led to a twelve-year centre of excellence funding grant, corpus linguistics took a longer time to spread to other language departments. After appointments in Helsinki and Tampere, Kytö moved to Uppsala in 1995. There she is continuing her work with corpus linguistics, although acquiring funding for the creation of new corpora is not always easy.

In terms of the second generation of corpus linguists dealing with spoken language our first case was *Åke Viberg* (b. 1945), who started out as a generative linguist and who even published a Swedish textbook on Chomsky's ideas. Over time he grew interested in natural languages and turned to corpora, a change that was facilitated by technological developments. His corpora have been used for studies of second-language acquisition, which has made it possible to finance the research from unconventional sources like the Swedish National Agency for Education. This was particularly

advantageous, since his research orientation initially met with resistance among colleagues in linguistics. Nowadays, the area is also funded by traditional research-funding bodies.

Our second example of a linguist dealing with spoken language, *Jens Allwood* (b. 1947), was an early sceptic of the ideas of Chomsky. Over the years he has moved between institutions and research interests. In terms of corpus linguistics, he has particularly followed Randolph Quirk, Jan Svartvik and others, who focused on spoken language. In so doing, he is particularly interested in the interaction between speakers and has gone even further than Quirk and his colleagues by including gestures and facial expressions. Another link back to the first generation is an educational programme at the University of Gothenburg in which he collaborated with Sture Allén.

It is evident that the conditions for the second generation were different from those of the four scholars belonging to the first generation. First of all, technological developments have made the creation of corpora much easier. Second, corpora appear to be much more readily accepted, indeed, even considered to be natural tools in linguistic research. This in turn may explain the fact that the second generation seems to face greater difficulty in financing the creation of new corpora. These are no longer seen as innovative as they were at the time of the first generation of innovators. Many corpora already exist, and there is a risk that funding bodies might ask what yet another corpus will add to our knowledge. It is therefore interesting to note that user-oriented corpora, such as those developed by Åke Viberg for second-language acquisition are considered more attractive for funding.

International perspectives

Since our research has pointed to the importance of considering the international context of research we have also undertaken an analysis of a database created from a search in SciVerse Scopus by using a profile with

words associated with corpus linguistics (Chapter 8). In this way we have identified significant works and significant authors. Our main findings from this search are as follows:

At the top of works published in the 1970s, 1980s and 1990s were three significant works in corpus linguistics. In the 1970s, it is *The American Heritage Word Frequency* (Carroll, Davis & Richman, 1971); in the 1980s, *A Comprehensive Grammar of the English Language* (Quirk et al., 1985); and in the 1990s, *The Longman Grammar of Spoken and Written English* (Biber et al., 1999).

The database of the most frequently cited authors includes two clusters containing two types of researchers: corpus builders and corpus users. The former is made up of linguists who create corpora, while the second consists of cognitive scientists who use corpora, focusing on language acquisition and language loss.

In terms of development over time, we can note that the above-mentioned three works are the central references in each of the three time clusters, and that the three time clusters are linked together by the classical British-born Australian linguist Michael A. K. Halliday and the Dutch psycholinguist Willem Levelt.

With one exception – Jan Svartvik, who worked closely with British colleagues – the Swedish corpus linguists do not appear among the top references. The same is also true for other non-Anglo-American corpus linguists like the French and the German, a circumstance which is an indication of an Anglo-American bias in the database.

In accordance with a model developed during the project we have also been able to show how the field has become increasingly organized over time. From the early 1960s onwards, a number of associations have been created both in Europe and North America, but also transnational ones. Several of these have also launched journals which have become important channels for the publishing of research in corpus linguistics. More recently this institutionalization has been manifested through the publication of handbooks and textbooks.

Concluding remarks

All in all, our study has shown that corpus linguistics, although such studies were undertaken already more than one hundred years ago, has particularly developed during the past fifty years. From being an innovation questioned by Chomskyans as well as some traditional linguists, corpora are nowadays standard tools in linguist research.

A final, and very important, conclusion of this study is that scientific innovations cannot be looked upon in a restricted national context. International developments in a field, including technological developments providing new methods and capabilities, are extremely important. At the same time, it is evident from our research that new national resources for the funding of innovative research are crucial in enabling individual actors to create a link to and develop the international research front in their home country.

For the future we can note that further technological advances will ensure both the accessibility of large corpora and opportunities to create small corpora. As a result, the divide between corpus linguists and generativists is no longer so dramatic. In addition, corpus linguistics has become key to modern information technology, for instance in smartphones. Corpus linguistics are therefore another demonstration of the unexpected use of basic research. Today the field enjoys wide use, both in academic work and in practice.

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ACH Association for Computing and the Humanities
ACL Association of Computational Linguistics
ADHO Alliance of Digital Humanities Organizations
ALLC Association for Literary and Linguistic Computing
AMTCL Association for Machine Translation and Computational Linguistics
BEC Bose-Einstein Condensation
CAMET Computer Archive of Modern English Texts
CL Corpus Linguistics
CLARIN Common Language Resources and Technology Infrastructure
CNRS Centre national de la recherche scientifique
COSTO Corpus of Stockholm
ECI/MCI European Corpus Initiative Multilingual Corpus
EURALEX European Association for Lexicography
EUROCORES EUROpean COllaborative RESearch
Evo-Devo Evolutionary Developmental Biology
FAS The Research Council for Working Life and Social Science (Forskningsrådet för arbetsliv och socialvetenskap)
FORMAS The Swedish Research Council for Sustainable Development (Forskningsrådet för miljö, areella näringar och samhällsbyggande)
FPM Français parlé des médias
FRN The Swedish Council for Planning and Co-ordination of Research (Forskningsrådsnämnden)
FUMS Forskning och Utbildning i Modern Svenska (Uppsala Universitet)
IAMT International Association for Machine Translation
ICAME International Computer Archive of Modern and Medieval English
ICCL International Committee on Computational Linguistics

IDS Institut für Deutsche Sprache
ILSA International Large Scale Student Assessments
INaLF Institut National de la Langue Française
KTH Royal Institute of Technology (Kungliga Tekniska Högskolan)
KWIC Key Words in Context
LDC Linguistic Data Consortium
LOB Lancaster-Oslo/Bergen Corpus
NAL Nordic Association of Linguists
NLP Natural Language Processing
OBC Old Bailey Corpus
OUP Oxford University Press
RFI The Swedish Research Council's Council for Research Infrastructure (Rådet för forskningens infrastrukturer)
RHESI Re-Structuring Higher Education and Scientific Innovation
SDH-SEMI Canadian Society for Digital Humanities
SEU Survey of English Usage
SIDA The Swedish International Development Cooperation Agency
SOU Swedish Government Official Reports (Statens Offentliga Utredningar)
SSE Survey of Spoken English
SSKKII Språk, Semantik, Kognition, Kommunikation, Interaktion och Information
SSM Swedish as Target Language (Svenska som målspråk)
TEFL Teaching English as a Foreign Language
TLFi Le Trésor de la Langue Française Informatisé
UCDL Uppsala Data Centre, Computational Linguistics
UDAC Uppsala Data Centre
VINNOVA Sweden's Innovation Agency (Verket för innovationssystem)
VR The Swedish Research Council (Vetenskapsrådet)

REFERENCES

Aarts, Jan & Willem Meijs (eds), 1984, *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*. Amsterdam: Rodopi.

Abelin, Åsa & Sture Allén, 1986, *Svensk ordbok* ('Swedish Dictionary'). Solna: Esselte Studium.

Aijmer, Karin & Bengt Altenberg (eds), 1991, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Allén, Sture, 1964, 'Ordforskaren och datamaskinen' ("The Linguist and the Computer"). *Göteborgs Handels- och Sjöfarts-Tidning*, March 31.

Allén, Sture, 1965, *Grafematisk analys som grundval för textedering: med särskild hänsyn till Johan Ekeblads brev till brodern Claes Ekeblad 1639–1655*. ('Graphemic Analysis as a Basis for Text Editing with Special Consideration of Johan Ekeblad's Letters to his Brother Claes Ekeblad 1639–1655'). Nordistica Gothoburgensia, 1 (diss.). Gothenburg: Almqvist & Wiksell.

Allén, Sture, 1970–1980, *Nusvensk frekvensordbok baserad på tidningstext: Frequency Dictionary of Present-day Swedish Based on Newspaper Material*. Data linguistica, 1-4. Stockholm: Almqvist & Wiksell International.

Allén, Sture, 1972, *Tiotusen i topp: ordfrekvenser i tidningstext* ('Ten Thousand in Top: Word Frequencies in Newspaper Texts'). Data linguistica, 6. Stockholm: Almqvist & Wiksell.

Allén, Sture, 1999, *Modersmålet i fäderneslandet* ('The Native Tongue in the Homeland'). Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Allwood, Jens, 1969, 'Medfödda idéer i Chomskys generativa grammatik' ("Native Ideas in Chomsky's Generative Grammar"). Unpublished paper. Gothenburg: University of Gothenburg, Department of Philosophy.

Allwood, Jens, 1976, *Linguistic Communication as Action and Cooperation*. Gothenburg Monographs in Linguistics, 2. Gothenburg: University of Gothenburg, Department of Linguistics (diss.).

Allwood, Jens, 2008a, 'Multimodal Corpora', in: Lüdeling, Anke & Merja Kytö (eds), *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, pp. 207–225.

Allwood, Jens, 2008b, 'Dimensions of Embodied Communication: Towards a Typology of Embodied Communication', in: Wachsmuth, Ipke, Manuela Lenzen & Günther Knoblich (eds), *Embodied Communication in Humans and Machines*. Oxford: Oxford University Press, Chapter 12.

Allwood, Jens, 2013, 'A Framework for Studying Human Multimodal Communication', in: Matej, Rojc & Nick Campbell (eds) *Coverbal Synchrony in Human-Machine Interaction*. Boca Raton, FL: CRC Press, Taylor & Francis Group, pp. 17–39.

Atkins, B. T. Sue & Antonio Zampolli, 1994, *Computational Approaches to the Lexicon*. Oxford: Oxford University Press.

Bauer, Marianne (ed.), 1999, *Transforming Universities. Changing Patterns of Governance, Structure and Learning in Swedish Higher Education*. Higher Education Policy Series, 48. London: Jessica Kingsley Publishers Ltd.

Benveniste, Émile, 1948, *Noms d'agent et noms d'action en indo-européen*. Paris: Adrien-Maisonneuve.

Benveniste, Émile, 1969, *Le vocabulaire des institutions indo-européennes, 1: Économie, parenté, société*. Paris: Les éditions de Minuit.

Berg, Sture, 1978, *Olika lika ord: svenskt homograflexikon* ('On Different Similar Words: Swedish Dictionary of Homographs'). Data linguistica, 12. Stockholm: Almqvist & Wiksell International.

Berko, Jean, 1958, 'The Child's Learning of English Morphology', *Word*, 14 (2–3), pp. 150–177.

Biber, Douglas, 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad & Randi Reppen, 1998, *Corpus Linguistics: Investigating Language. Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan, 1999, *The Longman Grammar of Spoken and Written English*. London: Longman.

Bloomfield, Leonard, 1933, *Language*. New York: Holt, Rinehart and Winston.

Borin, Lars, 1986, 'Is Hungarian a Case Language?', *Fenno-Ugrica Suecana*, 8, pp. 1–33.

Borin, Lars, 1991, *The Automatic Induction of Morphological Regularities*. Uppsala: Department of Linguistics.

Brattö, Olof, 1953, *Studi di antroponimia fiorentina: il libro di Montaperti (An. MCCLX)*. Gothenburg: University College of Gothenburg (diss.).

Brundenius, Claes, Bo Göransson & Jan Ågren, 2008, 'The Role of Academic Institutions in the National System of Innovation and Debate in Sweden', in: *UniDev Discussion Paper Series, Paper no 9*. Lund: Lund University Research Policy Institute.

Buchanan, Milton A., 1931, *A Graded Spanish Word Book*. Publications of The American and Canadian Committees on Modern Languages, 3. Toronto: University of Toronto Press.

Busa, Roberto, 1951, *Sancti Thomae Aquinatis Hymnorum ritualium varia specimina concordantiarum: Primo saggio di indici di parole automaticamente composti e stampati da machine IBM a schede perforate*. Milan: Bocca.

Bybee, Joan & Paul Hopper (eds), 2001, *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins Publishing.

Carroll, John B., Peter Davies & Barry Richman, 1971, *The American Heritage Word Frequency Book*. New York: Houghton Mifflin.

Chantraine, Pierre, 1933, *La formation des noms en grec ancien*. Paris: Champion.

Chantraine, Pierre, 1968, *Dictionnaire étymologique de la langue grecque*, 1. Paris: Éditions Klincksieck.
Cheydleur, Frederic D., 1934, *French Idiom List Based on a Count of 1,183,000 Running Words*. Publications of the American and Canadian Committees on Modern Languages, 16. New York: Macmillan.
Chomsky, Noam A., 1957, *Syntactic Structures*. New York: Mouton.
Chomsky, Noam A., 1965, *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
Chomsky, Noam A., 1966, *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. New York: Harper & Row.
Chomsky, Noam A. & Morris Halle, 1968, *The Sound Pattern of English*. New York: Harper & Row.
CNRS, 1961, *Colloque international, Strasbourg, 12–16 novembre, 1957: Lexicologie et lexicographie françaises et romanes: orientations et exigences actuelles*. Paris: Centre national de la recherche scientifique.
Coltheart, Max, Eddy Davelaar, Jon Torfi Jonasson & Derek Besner, 1977, 'Access to the Internal Lexicon', in: Dornic, Stan (ed.), *Attention and Performance,* VI. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 535–555.
Dahlberg, Carina, Pehr Hedenqvist & Ingrid Sundström (eds), 2017, *Knut and Alice Wallenberg Foundation*. Stockholm: Max Ström.
Danell, Karl-Johan, 1990, 'Corpus de journaux francophones sur ordinateur', *Travaux de linguistique*, 20, pp. 73–82.
Delcourt, Christian & Marc Hug (eds), 2009, *Mélanges offerts à Charles Muller pour son centième anniversaire (22 septembre 2009)*. Paris: CILF.
Dell, Gary S., 1986, 'A Spreading-activation Theory of Retrieval in Sentence Production', *Psychological Review*, 93 (3), pp. 283–321.
Durand, Jacques (ed.), 1983, *A Festschrift for Peter Wexler: Studies Presented to Peter Wexler by His Friends and Colleagues on the Occasion of His Sixtieth Birthday*. Occasional papers, 27. University of Essex. Colchester: Department of Language and Linguistics.
Eggers, Hans, 1969, (ed.), *Elektronische Syntaxanalyse der deutschen Gegenwartssprache: ein Bericht*. Tübingen: Niemeyer.

Elert, Claes-Christian, 1964, *Phonologic Studies of Quantity in Swedish: Based on Material from Stockholm Speakers*. Stockholm: Almqvist & Wiksell (diss.).

Ellegård, Alvar, 1953, *The Auxiliary Do: The Establishment and Regulation of its Use in English*. Gothenburg Studies in English. Stockholm: Almqvist & Wiksell (diss.).

Ellegård, Alvar, 1962, *A Statistical Method for Determining Authorship: the Junius Letters, 1769–1772*. Gothenburg Studies in English, 13. Stockholm: Almqvist & Wiksell.

Ellegård, Alvar, 1978, *The Syntactic Structure of English Texts: A Computer-based Study of Four Kinds of Text in the Brown University Corpus*. Gothenburg: Acta Universitatis Gothoburgensis, 43.

Elzinga, Aant, 1985, 'Research Bureaucracy and the Drift of Epistemic Criteria', in: Wittrock, Björn & Aant Elzinga (eds), *The University Research System: The Public Policies of the Home of Scientists*. Stockholm: Almqvist & Wiksell International, pp. 191–220.

Engwall, Gunnel, 1974, *Fréquence et distribution du vocabulaire dans un choix de roman français*. Stockholm: Skriptor språkförlag (diss.).

Engwall, Gunnel, 1978, 'Contenu, vocabulaire et statistique: illustration de quelques méthodes quantitatives', *Cahiers de lexicologie*, 33 (2), pp. 71–90.

Engwall, Gunnel, 1980, 'Le Plaidoyer d'un fou: Un plaidoyer de Strindberg ou de Loiseau?', *Stockholm Studies in Modern Philology, New Series, 6*. Stockholm: Almqvist & Wiksell International, pp. 29–54.

Engwall, Gunnel, 1984, *Vocabulaire du roman français (1962–1968): dictionnaire des fréquences*. Data linguistica, 17. Stockholm: Almqvist & Wiksell International.

Engwall, Gunnel, 1990, 'Les Vivisections de Strindberg dans la presse française de 1894 et 1895', in: Lindvall, Lars (ed.), *Actes du Xe congrès des romanistes scandinaves. Études Romanes de Lund 45*. Lund: Lund University Press, pp. 115–121.

Engwall, Gunnel, 1994a, 'Not Chance but Choice: Criteria in Corpus Construction', in: Atkins, Sue B. T. & Antonio Zampolli (eds), *Com-*

putational Approaches to the Lexicon. Automating the Lexicon, II. Oxford: Oxford University Press, pp. 49–82.
Engwall, Gunnel, 1994b, (ed.), *Strindberg et la France*. Romanica Stockholmiensia, 15. Stockholm: Acta Universitatis Stockholmiensis.
Engwall, Gunnel, 1995, 'Les formes verbales en suédois et en français: définitions et terminologies', *Travaux de linguistique*, 31, pp. 119-130.
Engwall, Gunnel, 1996, 'Corpus de français établis en Suède', *Revue Française de Linguistique Appliquée*, 1 (2), pp. 89-90.
Engwall, Gunnel, 1998, 'Strindberg et ses contacts français', in: Mellet, Sylvie & Marcel Vuillaume (eds), *Mots chiffrés et déchiffrés: Mélanges offerts à Étienne Brunet*. Paris: Honoré Champion Éditeur, pp. 473-501.
Engwall, Gunnel, 2009, 'Orthonet et les Vivisections d'August Strindberg', in: Delcourt, Christian & Marc Hug (eds), *Mélanges offerts à Charles Muller pour son centième anniversaire (22 septembre 2009)*. Paris: CILF, pp. 169-181.
Engwall, Gunnel & Inge Bartning, 1989, 'Le COSTO – description d'un corpus journalistique', *Moderna språk*, 83 (4), pp. 343-348.
Engwall, Lars, 1987, 'An American Dream. Postgraduate Research Training in the Social Sciences in Sweden', in: *Postgraduate Research Training in the Social Sciences*. Copenhagen: International Federation of Social Science Organizations, pp. 122-128.
Engwall, Lars, 2016, *Universitet under uppsikt* ('Universities under Supervision'). Stockholm: Dialogos.
Engwall, Lars, 2018, 'The Legacy of Knut and Alice: Governance of and by a Major Swedish Foundation', Manuscript. Department of Business Studies, Uppsala University.
Engwall, Lars & Thorsten Nybom, 2007, 'The Visible Hand Versus the Invisible Hand. The Allocation of Research Resources in Swedish Universities', in: Whitley, Richard & Jochen Gläser (eds), *The Changing Governance of the Sciences. The Advent of Research Evaluation Systems*. Berlin: Springer, pp. 31-49.

Engwall, Lars & Tina Hedmo, 2016, 'The Organizing of Scientific Fields: The Case of Corpus Linguistics', *European Review*, 24 (4), pp. 568-591.

Engwall, Lars, Enno Aljets, Tina Hedmo, Elias Håkansson & Raphaël Ramuz, 2015, 'Computer Corpus Linguistics: An Innovation in the Humanities', in: Whitley, Richard & Jochen Gläser (eds), *Organizational Transformation and Scientific Change: The Impact of Institutional Restructuring on Universities and Intellectual Innovation*. Research in the Sociology of Organizations, 42. Bingely: Emerald, pp. 331-365.

Enkvist, Erik, Charles A. Ferguson, Eva Hajičová & Peter Ladefoged, 1992, *Linguistic Research in Sweden*. Stockholm: The Swedish Council for Research in the Humanities and Social Sciences.

Ernout, Alfred & Antoine Meillet, 1932, *Dictionnaire étymologique de la langue latine: histoire des mots*. Paris: Klincksieck.

Etcherelli, Claire, 1967, *Élise ou la vraie vie*. Paris: Denoël.

Facchinetti, Roberta (ed.), 2007, *Corpus Linguistics 25 Years On*. Amsterdam: Rodopi.

Fillmore, Charles, 1992, '"Corpus Linguistics" or "Computer-aided Armchair Linguistics"', in: Svartvik, Jan (ed.), *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin: Mouton de Gruyter, pp. 35-60.

Firth, John Rupert, 1957, *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Forsgren, Mats, 2002, 'Le français parlé des médias (FPM): programme pour une recherche variationniste pluri-dimensionnelle', in: Dörum, Halvar, Marianne Hobaek Haff, Leif Sletsjö & Birte Stensgaard, (eds), *Actes du XVe congrès des romanistes scandinaves, Oslo 12–15 août 2002*. Oslo: Le Département des études classiques et romanesques, pp. 351–358.

Forster, Ken I., 1976, 'Accessing the Mental Lexicon', in: Wales, Roger J. & Edvard Walker (eds), *New Approaches to Language Mechanisms*. Amsterdam: North-Holland, pp. 257-287.

Francis, Nelson W. & Henry Kučera, 1982, *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Gesprochene Sprache, 1974. Düsseldorf: Schwann.

Gläser, Jochen, Enno Aljets, Adriane Gorga, Tina Hedmo, Elias Håkansson & Grit Laudel, 2015, 'Path Dependence and Policy Steering in the Social Sciences: The Varied Impact of International Large Scale Student Assessment on the Educational Sciences in Four European Countries', in: Whitley, Richard & Jochen Gläser (eds), *Organizational Transformation and Scientific Change: The Impact of Institutional Restructuring on Universities and Intellectual Innovation*. Research in the Sociology of Organizations, 42. Bingely: Emerald, pp. 267–295.

Grund, Peter J., 2004, *'Misticall Wordes and Names Infinite': An Edition of Humfrey Lock's Treatise on Alchemy. With an Introduction, Explanatory Notes and Glossary*. Uppsala: Department of English (diss.).

Grund, Peter J., 2011, *Misticall Wordes and Names Infinite: An Edition and Study of Humfrey Lock's Treatise on Alchemy*. Tempe, AZ: Arizona Center for Medieval and Renaissance Studies.

Gustavsson, Sverker, 1989, 'Den sista forskningspropositionen' ("The Last Research Bill"), in: Nybom, Thorsten (ed.), *Universitet och samhälle: Om forskningspolitik och vetenskapens samhälleliga roll* ('University and Society. On Research Policy and the Role of Science in Society'). Stockholm: Tiden, pp. 165–178.

Halliday, Michael A. K., 2004, *Lexicology and Corpus Linguistics: An Introduction*. London: Continuum.

Halliday, Michael A. K. & Ruqaiya Hasan, 1976, *Cohesion in English*. London: Longman.

Hammarberg, Björn & Åke Viberg, 1977, *Felanalys och språktypologi: orientering om två delstudier i SSM-projektet* ('Error Analysis and Language Typology: Orientation on Two Sub-Studies in the SSM Project'). Stockholm: Stockholm University (SSM report, 99-0277062-2).

Hebb, Donald Olding, 1949, *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley.

Henmon, Vivian A. C., 1924, *A French Word Book Based on a Count of 400 000 Running Words*. Bureau of Educational Research Bulletins, 3. Madison: University of Wisconsin Press.

Hinc robur et securitas: en forskningsstiftelses handel och vandel: Stiftelsen Riksbankens jubileumsfond 1989–2003, 2004, ('Hinc robur et securitas: the Conduct of a Research Foundation: The Bank of Sweden Tercentenary Foundation 1989–2003'). Hedemora: Gidlunds.

Högskoleförordningen 1977:263 ('Ordinance for Universities and University Colleges'). Stockholm: Utbildningsdepartementet.

Högskoleverket, 2007, *Befordran till professor och lektor: en rättslig översikt* ('Promotion to Professor or Lecturer: A Legal Overview'). Rapport 2007:55 R. Stockholm: Högskoleverket.

Hoppe, Gunnar, Gert Nylander & Ulf Olsson, 1993, *Till landets gagn: Knut och Alice Wallenbergs stiftelse 1917–1992* ('For the Benefit of the Nation: the Knut and Alice Wallenberg Foundation 1917–1992'). Stockholm: Norstedts.

http://icame.uib.no/history/founding_document_1977.pdf (Accessed on July 28, 2017).

http://spraakbanken.gu.se/ (Accessed on July 29, 2017).

http://www.aclweb.org/archive/misc/History.html (Accessed on July 28, 2017).

http://www.guardian.co.uk/higher-education-network/blog/2011/aug/12/father-roberto-busa-academic-impact (Accessed on July 28, 2017).

http://www.mt-archive.info/LREC-2004-Zampolli.pdf (Accessed on July 28, 2017).

http://www.natcorp.ox.ac.uk/ (Accessed on July 28, 2017).

http://www.su.se/profiles/bartn-1.195116 (Accessed on July 29, 2017).

http://www.svenskaakademien.se/svenska-akademien/de-aderton/stol-nr-3-sture-allen (Accessed on July 29, 2017).

https://ki.se/nyheter/sa-gar-det-till-att-bli-professor-pa-ki (Accessed on February 15, 2018).

https://spraakbanken.gu.se/swe/forskning/infrastruktur/swe-clarin (Accessed on February 19, 2018).
https://www.elsevier.com/solutions/scopus (Accessed on July 28, 2017).
Huddleston, Rodney & Geoffrey K. Pullum, 2002, *Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Hunston, Susan & Gill Francis, 2000, *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Philadelphia: John Benjamins.
Isaksson, Bo & Håkan Malmberg, 2005, 'The Establishment of the Chair of Semitic Languages and a Few Words on its History', in: Nordesjö, Hans (ed.), *The Call of the Orient: The Professorship of Semitic Languages at Uppsala University 400 Years*. Uppsala: Uppsala Universitetsbibliotek, pp. 5–11.
Jakobson, Roman, 1941, *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala.
Jakobson, Roman C., Gunnar M. Fant & Morris Halle, 1961, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press.
Johansson, Stig, 2008, 'Some Aspects of the Development of Corpus Linguistics in the 1970s and 1980s', in: Lüdeling, Anke & Merja Kytö (eds), *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, pp. 33–53.
Jonsson, Inge, 2003, *Vitterhetsakademien 1753–2003* ('The Royal Swedish Academy of Letters, History and Antiquities 1753–2003'). Stockholm: Kungliga Vitterhets Historie och Antikvitets Akademien.
Juilland, Alphonse & Eugenio Chang-Rodriguez, 1964, *Frequency Dictionary of Spanish Words*. The Romance Languages and their Structures. First series, S1. The Hague: Mouton.
Juilland, Alphonse, Dorothy Brodin & Catherine Davidovitch, 1970, *Frequency Dictionary of French Words*. The Romance Languages and their Structures. First series, F1. The Hague: Mouton.

Juilland, Alphonse, Vincenzo Paolo Traversa & Antonio Beltramo, 1973, *Frequency Dictionary of Italian Words*. The Romance Languages and their Structures. First series, I1. The Hague: Mouton.

Juilland, Alphonse, P. M. H. Edwards & Ileana Juilland, 1965, *Frequency Dictionary of Rumanian Words*. The Romance Languages and their Structures. First series, R1. The Hague: Mouton.

Kaeding, Fredrich Wilhelm, 1897–1898, *Häufigkeitswörterbuch der deutschen Sprache*, 1–2. Steiglitz bei Berlin: Selbstverlag des Herausgebers.

Kučera, Henry & Nelson W. Francis, 1967, *Computational Analysis of Present-Day American English*. Providence, RI: Providence University Press.

Kuhn, Thomas S., 1962, *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kytö, Merja, 1991, *Variation and Diachrony, with Early American English in Focus: Studies on CAN/MAY and SHALL/WILL*. Bamberger Beiträge zur englischen Sprachwissenschaft. Frankfurt am Main: Lang.

Kytö, Merja, 1996, 3rd ed., *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*. Helsinki: Department of English, University of Helsinki

Kytö, Merja, Peter J. Grund & Terry Walker, 2011, *Testifying to Language and Life in Early Modern England. Including a CD-ROM containing an Electronic Text Edition of Depositions 1560–1760* (ETED). Amsterdam: John Benjamins.

Labov, William, 1972, *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Landberg, Hans, Olle Edqvist & Uno Svedin, 1995, *Parliament, Government and Research. Some Features of Swedish Research Policy over Two Decades*. Report of the Commission to Review the Structure for Finance Research, Stockholm: Swedish Council for Planning and Coordination of Research.

Laudel, Grit, Eric Lettkemann, Raphaël Ramuz, Linda Wedlin & Richard Woolley, 2015a, 'Cold Atoms – Hot Research: High Risks, High

Rewards in Five Different Authority Structures', in: Whitley, Richard & Jochen Gläser (eds), *Organizational Transformation and Scientific Change: The Impact of Institutional Restructuring on Universities and Intellectual Innovation*. Research in the Sociology of Organizations, 42. Bingely: Emerald, pp. 203–234.

Laudel, Grit, Martin Benninghoff, Eric Lettkemann & Elias Håkansson, 2015b, 'Highly Adaptable but not Invulnerable: Necessary and Facilitating Conditions for Research in Evolutionary Developmental Biology', in: Whitley, Richard & Jochen Gläser (eds), *Organizational Transformation and Scientific Change: The Impact of Institutional Restructuring on Universities and Intellectual Innovation*. Research in the Sociology of Organizations, 42. Bingely: Emerald, pp. 235–265.

Le Clézio, Jean-Marie Gustave, 1966, *Le Déluge*, Paris: Gallimard.

Leech, Geoffrey, Paul Rayson & Andrew Wilson, 2001, *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.

Levelt, Willem J. M., 1989, *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt, Willem J. M. & Linda Wheeldon, 1994, 'Do Speakers have Access to a Mental Syllabary?', *Cognition*, 50 (1–3), pp. 239–269.

Levelt, Willem J. M., Ardi Roelofs & Antje S. Meyer, 1999, 'A Theory of Lexical Access in Speech Production', *Behavioral and Brain Sciences*, 22 (1), pp. 1–75.

Lindblom, Björn, 1968, *On the Production and Recognition of Vowels*. Lund: Lund University (diss.).

Lüdeling, Anke & Merja Kytö (eds), 2008, *Corpus Linguistics: An International Handbook*, Vol. 1. Berlin: Mouton de Gruyter.

Lüdeling, Anke & Merja Kytö (eds), 2009, *Corpus Linguistics: An International Handbook*, Vol. 2. Berlin: Mouton de Gruyter.

Malmberg, Bertil, 1954, *La phonétique*. Paris: Presses universitaires de France.

Malmberg, Bertil, 1963, *Structural Linguistics and Human Communication: An Introduction into the Mechanism of Language and the Methodology of Linguistics*. Berlin: Springer.
Malmberg, Bertil, 1966, *Les nouvelles tendances de la linguistique*. (Translation by Jacques Gengoux of *Nya vägar inom språkvetenskapen*.) Paris: Presses universitaires de France.
Malmberg, Bertil, 1970, *Nouvelles perspectives en phonétique*. Bruxelles: Presses universitaires de Bruxelles.
Malmberg, Bertil, 1977, *Signes et symboles: les bases du langage humain*. Paris: Picard.
McCarthy, Michael & Anne O'Keefe, 2010, 'Historical Perspective: What are Corpora and How Have They Evolved?', in: O'Keeffe, Anne & Michael McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, pp. 3–13.
McClelland, James L., David E. Rumelhart & the PDP research group, 1986, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, II. Cambridge, MA: MIT Press.
McEnery, Tony & Andrew Hardie, 2011, *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
McEnery, Tony & Andrew Wilson, 1996, 2nd ed., 2001, *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
McEnery, Tony, Richard Xiao & Yukio Tono, 2006, *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.
Meurman-Solin, Anneli, 1993, *Variation and Change in Early Scottish Prose: Studies Based on the Helsinki Corpus of Older Scots*. Helsinki: Suomalainen Tiedeakatemia (diss.).
Meyer, Charles F., 2002, *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
Miller, George A., 1956, 'The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information', *Psychological Review*, 63 (2), pp. 81–97.
Morgan, Bayard Quincy, 1933, *German Frequency Word Book, Based on Kaeding's Häufigkeitswörterbuch der deutschen Sprache*. Publications of

the American and Canadian Committees on Modern Languages, 9. New York: Macmillan.
Morton, John, 1969, 'Interaction of Information in Word Recognition', *Psychological Review*, 76 (2), pp. 165–178.
Muller, Charles, 1967, *Étude de statistique lexicale: le vocabulaire du théâtre de Pierre Corneille*. Université de Strasbourg. Paris: Larousse (diss.).
Muller, Charles, 1968, *Initiation à la statistique linguistique*. Paris: Larousse.
Muller, Charles, 1979, *Langue française et linguistique quantitative: recueil d'articles*. Geneva: Slatkine.
Nationalencyklopedins ordbok ('The Dictionary of the National Encyclopedia'), 1995–1996. Höganäs: Bra böcker.
Nyberg, H. S. (ed.), 1943, *Orientering i språkvetenskap* ('Orientation in Language Studies'). Stockholm: Natur och kultur.
Nybom, Thorsten, 1997, *Kunskap, politik, samhälle: essäer om kunskapssyn, universitet och forskningspolitik 1900–2000* ('Knowledge, Politics, Society: Essays on Epistemological Approach, Universities and Research Policy 1900–2000'). Hargshamn: Arete.
Ochs, Elinor, Emanuel Schegloff & Sandra Thompson, 1996, *Interaction and Grammar*. Cambridge: Cambridge University Press.
O'Keeffe, Anne & Michael McCarthy (eds), 2010, *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge.
Öhman, Sven, 1968, *Studies in Articulatory Coordination*. Stockholm: Royal Institute of Technology (diss.).
Öhman, Sven, 2007, *The Essence of Language: A Philosophical Problem: Why Noam Chomsky was Never a Linguist*. Nora: Nya Doxa.
Öhrström, Lilian, 1991, *Research: the Swedish Approach*. Stockholm: Swedish Institute.
Oldfield, Richard C. & Arthur Wingfield, 1965, 'Response Latencies in Naming Objects', *Quarterly Journal of Experimental Psychology*, 17 (4), pp. 273–281.
Olsson, Leif-Jöran & Lars Borin, 2000, 'A Web-based Tool for Exploring Translation Equivalents on Word and Sentence Level in Multi-

lingual Parallel Corpora', in: *Erikoiskielet ja käännösteoria. VAKKI:n julkaisut*, N:o 26. Vaasa, pp. 76-84.
Peralta, Julia, 2008, 'Från matematikmaskin till IT: Datorhistoria på universitetsområdet' ("From Mathematical Machine to IT: Computer History in Universities"). Interview November 4, 2008 (mimeo).
Perec, George, 1965, *Les Choses: une histoire des années soixante*. Paris: Julliard.
Pierrehumbert, Janet B., 2001, 'Exemplar Dynamics: Work Frequency, Lenition and Contrast', in: Bybee, Joan & Paul Hopper (eds), 2001, *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins, pp. 137-157.
Pokorny, Julius von, 1959, *Indogermanisches etymologisches Wörterbuch*, Bd 1. Bern: Francke.
Premfors, Rune, 1986, *Svensk forskningspolitik* ('Swedish Research Policy'). Lund: Studentlitteratur.
Quemada, Bernard (ed.), 1959-1993, *Matériaux pour l'histoire du vocabulaire français: datations et documents lexicographiques, Part 1-30*. Besançon: Centre d'étude du vocabulaire français.
Quemada, Bernard, 1968, *Les dictionnaires du français moderne 1539-1863: étude sur leur histoire, leurs types et leurs méthodes*. Paris: Didier (diss.).
Quirk, Randolph, 1959, 'Relative Clauses in Educated Spoken English', *English Studies*, 38 (1), pp. 97-109.
Quirk, Randolph & Jan Svartvik, 1966, *Investigating Linguistic Acceptability*. The Hague: Mouton.
Quirk, Randolph & Jan Svartvik, 1972, *A Grammar of Contemporary English*. London: Longman.
Quirk, Randolph & Jan Svartvik, 1978, *A Corpus of Modern English*. Lund: University of Lund.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik, 1985, *A Comprehensive Grammar of the English Language*. London: Longman.
Regeringens proposition 1991/92:92, *Om utskiftning av löntagarfondernas tillgångar m.m.* ('The Bill 1991/92:92 of the Government: On the Dis-

tribution of the Assets of the Wage Earners Fund, etc.'). Stockholm: Fritzes.
Regeringens proposition 2000/01:3, *Forskning och förnyelse* ('The Bill 2000/01:3 of the Government: Research and Renewal'). Stockholm: Utbildningsdepartementet.
Renouf, Antoinette & Andrew Enouf, 2009, *Corpus Linguistics: Refinements and Reassessments*. Amsterdam: Rodopi.
Rissanen, Matti, 1967, *The Uses of One in Old and Early Middle English*. Mémoires de la Société néophilologique de Helsinki. Helsinki: University of Helsinki (diss.).
Rissanen, Matti, Merja Kytö & Minna Palander-Collin (eds), 1993, *Early English in the Computer Age: Explorations through the Helsinki Corpus*. Berlin: de Gruyter.
Rissanen, Matti, Merja Kytö & Kirsi Heikkonen (eds), 1997a, *English in Transition: Corpus-based Studies in Linguistic Variation and Genre Styles*. Berlin: de Gruyter.
Rissanen, Matti, Merja Kytö & Kirsi Heikkonen (eds), 1997b, *Grammaticalization at Work: Studies of Long-term Developments in English*. Berlin: de Gruyter.
Rosengren, Inger, 1966, *Semantische Strukturen: eine quantitative Distributionsanalyse einiger mittelhochdeutscher Adjektive*. Lunder germanistische Forschungen, 38. Lund: Gleerup (diss.).
Rosengren, Inger, 1972, *Ein Frequenzwörterbuch der deutschen Zeitungssprache: Die Welt, Süddeutsche Zeitung*, 1. Lunder germanistische Forschungen, 41. Lund: LiberLäromedel/Gleerup.
Rosengren, Inger, 1977, *Ein Frequenzwörterbuch der deutschen Zeitungssprache: Die Welt, Süddeutsche Zeitung*, 2. Lunder germanistische Forschungen, 43. Lund: LiberLäromedel/Gleerup.
Rosengren, Inger (ed.), 1981, *Sprache und Pragmatik: Lunder Symposium 1980*. Lunder germanistische Forschungen, 50. Lund: LiberLäromedel/Gleerup.

Rosengren, Inger (ed.), 1984, *Sprache und Pragmatik: Lunder Symposium 1984*. Lunder germanistische Forschungen, 53. Stockholm: Almqvist & Wiksell International.

Rosengren, Inger (ed.), 1986, *Sprache und Pragmatik: Lunder Symposium 1986*. Lunder germanistische Forschungen, 55. Stockholm: Almqvist & Wiksell International.

Rundgren, Frithiof, 1978, 'Språkvetenskaplig forskning' ("Linguistic Research"), in: Strömholm, Stig, Torgny Nevéus & Åke Davidsson (eds), *Universitet i utveckling: Uppsala universitet under Torgny T. Segerstedts rektorat 1955–1978. Del II. Forskningen vid Uppsala Universitet 1955–1978. Universitetsbibliotekets utveckling* ('A University in Development: Uppsala University During the Vice-Chancellorship of Torgny T. Segerstedt 1955-1978. Part II. The Research at Uppsala University 1955–1978. The Development of the University Library'). Skrifter rörande Uppsala universitet, C, Organisation och historia. Uppsala: Acta Universitatis Upsaliensis, pp. 97–102.

Sagan, Françoise, 1965, *La Chamade*. Paris: Julliard.

Saussure, Ferdinand de, Charles Bally & Albert Sechehaye, 1916, *Cours de linguistique générale*. Lausanne: Payot.

Schegloff, Emanuel A., 2006, *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge: Cambridge University Press.

Schwyzer, Eduard, 1939, *Griechische Grammatik*, Bd 1. *Allgemeiner Teil, Lautlehre, Wortbildung, Flexion*. München: Beck'sche Vlgs-Buchhandlung.

Seidenberg, Mark S. & James L. McClelland, 1989, 'A Distributed, Developmental Model of Word Recognition and Naming', *Psychological Review*, 96 (4), pp. 523–568.

Shannon, Claude, 1948, 'A Mathematical Theory of Communication', *The Bell System Technical Journal*, 27 (3 & 4), pp. 379–423; 623–656.

Sinclair, John McHardy, 1987, *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London: HarperCollins.

Sinclair, John McHardy, 1991, *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Skoie, Hans, 2001, 'The Research Councils in the Nordic Countries: Development and Some Challenges', in: *Report 10/2001*. Oslo: NIFU – Norsk Institutt for studier av forskning og utdanning.

Sörlin, Sverker, 2005, *Vad kan stiftelser göra?: den privata stiftelsesektorn som forskningsfinansiär* ('What Can Foundations Do: The Sector of Private Foundations as a Financier of Research'). Örnsköldsvik: Kempestiftelserna.

SOU 1975:26, *Forskningsråd: betänkande av Forskningsrådsutredningen* ('Research Councils: Report from the Research Council Committee'). Stockholm: Liber.

SOU 2016:29, *Trygghet och attraktivitet: en forskarkarriär för framtiden* ('Security and Attractiveness: A Research Career for the Future'). Stockholm: Wolter Kluwers.

Statistical Yearbook of Sweden 1956. Stockholm: Statistics Sweden.

Strategic Plan for the Development of Research in Language Technology at the University of Gothenburg, 2009. Gothenburg: University of Gothenburg.

Stubbs, Michael, 2002, *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Svartvik, Jan, 1966, *On Voice in the English Verb*. Uppsala University. The Hague: Mouton (diss.).

Svartvik, Jan (ed.), 1990, *The London-Lund Corpus of Spoken English: Description and Research*. Lund Studies in English, 82. Lund: Lund University Press.

Svartvik, Jan, 2005, 'A Life in Linguistics', *European English Messenger*, 14 (1), pp. 34–44.

Svartvik, Jan & Randolph Quirk (eds), 1980, *A Corpus of English Conversation*. Lund Studies in English, 56. Lund: LiberLäromedel/Gleerup.

Taft, Marcus, 1979, 'Recognition of Affixed Words and the Word Frequency Effect', *Memory and Cognition*, 7 (4), pp. 263–272.

Teleman, Ulf, 1969, *Studies in a Generative Grammar of Modern Swedish*. Lund: Lund University (diss.).
Teubert, Wolfgang & Anna Čermáková, 2007, *Corpus Linguistics: A Short Introduction*. London: Continuum.
Thomas, Jenny & Michael Short (eds), 1996, *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London: Longman.
Thorndike, Edward L. & Irving Lorge, 1944, *The Teacher's Word Book of 30,000 Words*. New York: Teacher's College, Columbia University.
Tilander, Gunnar, 1957, *Nouveaux essais d'étymologie cynégétique*. Stockholm: Cynegetica.
Tognini-Bonelli, Elena, 2001, *Corpus Linguistics at Work*. Studies in Corpus Linguistics, 6. Amsterdam: John Benjamins.
Trampe, Peter af & Åke Viberg, 1972, *Allmän språkteori och grammatik: en introduktion* ('General Language Theory and Grammar: An Introduction'). Lund: Gleerup.
Trésor de la langue française informatisé, 2004, *Analyse et traitement informatique de la langue française*. Centre national de la recherche scientifique, Université de Nancy II. Paris: CNRS editions.
Tunberg, Sven, 1957, *Stockholms högskolas historia före 1950* ('The History of Stockholm University College before 1950'). Stockholm: Norstedts.
Vander Beke, George E. 1929, *French Word Book*. Publications of the American and Canadian Committees on Modern Languages, 15. New York: Macmillan.
Vem är det 1997. Svensk biografisk handbok ('Who's Who 1997, Swedish Biographical Handbook'). Stockholm: Norstedts.
Vem är det 2007. Svensk biografisk handbok ('Who's Who 2007, Swedish Biographical Handbook'). Malmö: Nationalencyklopedin.
Viberg, Åke, 1981, *Studier i kontrastiv lexikologi: en typologisk och kontrastiv jämförelse av tre semantiska fält i svenskan: perceptionsverb, kognitiva predikat, emotiva predikat* ('Studies in Contrastive Lexicology: A Typological and Contrastive Comparison of Three Semantic Fields in Swedish: Verbs of Perception, Cognitive Predicates, Emotive Predicates'). Stockholm: Stockholm University (diss.).

Viberg, Åke, 1983, 'The Verbs of Perception: A Typological Study', *Linguistics*, 21 (1), pp. 123–162.

Walker, Terry, 2005, *Second Person Singular Pronouns in Early Modern English Dialogues*. Uppsala: Department of English, Uppsala University (diss.).

Walker, Terry, 2007, *Thou and You in Early Modern English Dialogues: Trials, Depositions, and Drama Comedy*. Amsterdam and Philadelphia: John Benjamins.

Wallander, Jan, 2002, *The Wenner-Gren Foundations 1955–2000: How Vanity, Visions and Over-ambitious Plans to Improve the World Led to the Creation of Great Foundations*. Stockholm: Atlantis.

West, Michael, 1953, *A General Service List of English Words with Semantic Frequencies and a Supplementary Word-list for the Writing of Popular Science and Technology*. London: Longman.

Whitley, Richard, 1984, *The Intellectual and Social Organization of the Sciences*. Oxford: Oxford University Press.

Whitley, Richard & Jochen Gläser (eds), 2015, *Organizational Transformation and Scientific Change: The Impact of Institutional Restructuring on Universities and Intellectual Innovation*. Research in the Sociology of Organizations, 42. Bingely: Emerald.

Williams, Geoffrey (ed.), 2005, *La linguistique de corpus*. Rennes: Presses universitaires de Rennes.

Windahl, Sven, 2013, 'Professor var enastående handledare' ("A Professor Being an Outstanding Supervisor"). *Sydsvenska Dagbladet*, July 30. https://www.sydsvenskan.se/2013-07-30/professor-var-enastaende-handledare. (Accessed on August 3, 2017.)

Yule, G. Udny, 1944, *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Zampolli, Antonio, Laura Cignoni & Carol Peters (eds), 1981, *Computational Lexicology and Lexicography. Special Issue Dedicated to Bernard Quemada*, I-II. Linguistica Computazionale VI and VII, Pisa: Giardini.

Zipf, George Kingsley, 1932, *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press.
Zipf, George Kingsley, 1935, *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Boston: MA: Houghton Mifflin Company.
Zipf, George Kingsley, 1949, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.

NAME INDEX

SUBJECT INDEX

I KVHAA *Filologisk-filosofiska serien* har följande arbeten utkommit:

1 Ström, F., *Diser, nornor, valkyrjor. Fruktbarhetskult och sakralt kungadöme i Norden* (Disen, Nornen, Walküren. Fruchtbarkeitskult und sakrales Königtum im Norden). 1954
2 Ekwall, E., *Studies on the population of Medieval London*. 1956
3 Kjellén, A., *Diktaren och havet. Drift- och drömsymbolik i svenskspråkig lyrik 1880–1940* (The poet and the sea). 1957
4 *Svenska skrock och signerier, samlade av Leonhard Fredrik Rääf* (Popular superstitions and incantations in Sweden collected by Leonhard Fredrik Rääf. Ed. with an introduction and textual notes by K. R. V. Wikman). 1957
5 Wessén, E., *Runstenen vid Röks kyrka* (Der Runenstein von Rök, Östergötland). 1958
6 Wessén, E., *Historiska runinskrifter*. 1960
7 Ståhl, H., *Ortnamnen i Kopparbergslagen*. 1960
8 Rooth, E., *Zu den Bezeichnungen für "Eiszapfen" in den Germanischen Sprachen. Historisch-wortgeografische und etymologische Studien*. 1961
9 Wessén, E., *Svensk medeltid. En samling uppsatser om svenska medeltidshandskrifter och texter. I. Landskapslagar*. (Zusammenfassung) 1968
10 Wessén, E., *Svensk medeltid. En samling uppsatser om svenska medeltidshandskrifter och texter. II. Birgittatexter*. (Zusammenfassung) 1968
11 Rooth, E., *Niederdeutsche Breviertexte des 14. Jahrhunderts aus Westfalen*. 1969
12 Dixelius, O., *Hans Järta och litteraturen. Hans Järta i litterär debatt och kulturpolitik under romantikens tidevarv 1809–1825* (Hans Järta and literature. Hans Järta in literary and cultural politics in the romantic era. 1809–1825). 1973. ISBN 91-7192-061-7
13 Beijer, A., *Dramatiken i Bröllops Besvärs Ihugkommelse. En tidsbild och ett tolkningsförsök*. 1974. ISBN 91-7192-173-7

14 Ridderstad, P., *Konsten att sätta punkt. Anteckningar om stenstilens historia 1400–1765*. 1975. ISBN 91-7192-242-3

15 *Proceedings of the VIth congress of Arabic and Islamic studies*. 1975. ISBN 91-7192-209-1

16 Wessén, E., *Svensk medeltid. En samling uppsatser om svenska medeltidshandskrifter och texter. III. De fornsvenska handskrifterna av Heliga Birgittas Uppenbarelser*. 1976. ISBN 91-7402-011-0

17 *Proceedings of the international colloquium on gnosticism*. Stockholm August 20–25 1973. 1977. ISBN 91-7402-025-0

18 Norberg, D., *Lœuvre poétique de Paulin d'Aquilée*. 1979. ISBN 91-7402-092-7

19 Sarajas, A., *Studiet av folkdiktningen i Finland intill slutet av 1700-talet*. 1982. ISBN 91-7402-144-3

20 Johns Blackwell, M., *C. J. L. Almqvist and Romantic Irony*. 1983. ISBN 91-7402-119-2

21 Makaev, È. A., *The language of the oldest runic inscriptions. A linguistic and historical analysis*. Translated from the Russian by J. Meredig in cooperation with E. H. Antonsen. 1996. ISBN 91-7402-259-8

22 Landgren, B., *Den hotade idyllen. Gunnar Mascoll Silfverstolpe, Finland och den lyriska intimismen*. 2008. ISBN 978-91-7402-379-4

23 Lindberg, B. (ed.), *The Pufendorf Lectures. Annotations from the teaching of Samuel Pufendorf 1672–1674*. 2014. ISBN 978-91-7402-426-5

24 Hidal, S., *Ivan Engnell. En bibelforskares bana*. 2019. ISBN 978-91-88763-01-3

25 Engwall, L., Hedmo, T. & Persson, O., *Corpus linguistics in Sweden: Pioneers and their context*. 2019. ISBN 978-91-88763-02-0