

János S. Petőfi (1931-2013) was one of the founders of Text Linguistics in Germany in the early '70s. He developed different text models, the most famous of which were the Text Structure World Structure Theory (TeSWeST) and Semiotic Textology. In this volume, some of his colleagues and disciples discuss his theoretical contributions to prove the enormous impact of his thoughts in the fields of linguistics, literary theory, rhetoric and semiotics. The essays here consider the notion of coherence, which Petőfi deemed to be the only sufficient condition for textuality, the relationships between his textual models and disciplines such as cognitive, computational and corpus linguistics, and his contributions to the analysis of literary and multimedial texts.

Margarita Borreguero Zuloaga studied Spanish and Italian Philology and Philosophy at the Complutense University of Madrid, where she is now Associate Professor at the Romance Studies Department. She has been a visiting scholar at the Universities of Heidelberg (with a Humboldt Fellowship), Bergamo, Marburg, Tübingen and Turin. Her research interests are in the fields of contrastive and text linguistics, with particular focus on information structure, discourse markers and anaphors. She is working on the acquisition of Italian and Spanish discourse markers and has authored more than 60 papers and the book *La interfaz lengua-texto. Un modelo de estructura informativa* (2015), in cooperation with Angela Ferrari. She is currently co-editor-in-chief of the journal *Cuadernos de Filología Italiana* and the book series *Sprache – Gesellschaft – Geschichte*.

Luciano Vitacolonna is a Researcher at the Department of Modern Languages, Literatures and Cultures of the University of Chieti-Pescara, Italy. His main research interests are in textual theory, literary theory and semiotics. His publications include *Per un paradigma semiotico del testo* (1989), *Principi e contributi di semiotica del testo* (1999), *Divagazioni testuali. Studi di testologia semiotica* (2000), *Semiotica* (2008), and *Studi di semiotica testuale* (2008). He edited, together with Petőfi, *Analisi e interpretazione dei testi letterari* (special issue of *Versus*, 1983) and the 3rd volume of the series *Sistemi segnici e il loro uso nella comunicazione umana* devoted to *La Testologia Semiotica e la comunicazione umana multimediale* (1996).

978-1-5275-2310-4
www.cambridgescholars.com
Cover image János S. Petőfi
© Vienna Tordone, 2015



Cambridge
Scholars
Publishing



The Legacy of János S. Petőfi

Margarita Borreguero Zuloaga
and Luciano Vitacolonna



The Legacy of János S. Petőfi

*Text Linguistics,
Literary Theory
and Semiotics*

Edited by
Margarita Borreguero Zuloaga
and Luciano Vitacolonna

The Legacy of János S. Petőfi

The Legacy of János S. Petőfi:

*Text Linguistics, Literary
Theory and Semiotics*

Edited by

Margarita Borreguero Zuloaga
and Luciano Vitacolonna

Cambridge
Scholars
Publishing



The Legacy of János S. Petőfi:
Text Linguistics, Literary Theory and Semiotics

Edited by Margarita Borreguero Zuloaga and Luciano Vitacolonna

This book first published 2019

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Copyright © 2019 by Margarita Borreguero Zuloaga, Luciano Vitacolonna
and contributors

All rights for this book reserved. No part of this book may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
electronic, mechanical, photocopying, recording or otherwise, without
the prior permission of the copyright owner.

ISBN (10): 1-5275-2310-1

ISBN (13): 978-1-5275-2310-4

CONTENTS

Introduction

The Lasting Legacy of János S. Petőfi viii
Margarita Borreguero

In Memory of János S. Petőfi xx
Luciano Vitacolonna

Linguistics

On the Text Organizing Status of Key Words in News Discourse:
An Empirical Study in Terms of János S. Petőfi's Legacy 2
József Andor

From Text Grammar to Macropragmatics via Coherence 24
Klaus Hölker

Petőfi's View of Language and Current Cognitive Linguistics 34
Enrique Bernárdez

Corpus-based Text Linguistics 53
Manuel Barbera, Carla Marelló

On a Computational Semiotic Textology 69
Borja Navarro Colorado

Literary Theory and Rhetoric

The Pragmatics in János S. Petőfi's Text Theory and the Cultural
Rhetoric: The Extensional-Semantic Code and the Literature
of the Spanish Golden Age 92
Tomás Albaladejo

János S. Petőfi's Linguistic and Textual Theory and the Recovery
of the Historical Thinking about Rhetoric 110
Francisco Chico Rico

Philology within the Framework of Semiotic Textology:
János S. Petőfi's Scientific Legacy II 132
Mauro Giuffrè

Semiotics

Semiotic-Textological interpretation of Communication in Art
Museum 168
Giuliana Pascucci

Appendix

On Basic, Disciplinary Issues of Textology and the Status of Semiotic
Textology: An Electronic Interview with János Sándor Petőfi..... 182
József Andor

CORPUS-BASED TEXT LINGUISTICS*

MANUEL BARBERA - CARLA MARELLO

ABSTRACT: Starting from the lesson Petőfi delivered in 2004 upon being awarded *Laurea honoris causa* from Turin University and from the Round Table that followed, our paper aims to trace the intertwined history of the relationship between corpus linguistics and text linguistics. Currently, that relationship is good and still improving, but that was not always the case: the role played by generative linguistics in opposing corpus linguistics is discussed in detail. The position of Petőfi, who was basically sceptical of the use of corpus linguistics, is traced back to this conflict. Finally we suggest that Petőfi's textology - as he described it - might be seen as an extension of text linguistics encompassing multimodal aspects.

KEYWORDS: Text linguistics; corpus linguistics; history of linguistics; Petőfi's textology

0. Antecedent

On the morning of February 27th, 2004, a round table was organized at the University of Turin, Italy, by a research group of the local Doctoral School in Linguistics, Applied Linguistics and Linguistic Engineering. The topic was *Linguistica del testo, semiotica del testo, linguistica dei corpora: quali rapporti? (Text Linguistics, Text Semiotics and Corpus Linguistics: Which Types of Connections?)*. János Sándor Petőfi was going to receive the *Laurea honoris causa* in Foreign Languages and Literatures in the afternoon¹ and was so kind to take part in the round table together

* The authors planned the content of this paper together; however, paragraphs 0, 1, 4 are by Carla Marello; 2, 3 and 5 by Manuel Barbera. James Alexander Ryan kindly helped with the English version.

¹ He offered as *Lectio Magistralis Il conferimento del senso ai testi verbali Aspetti dell'insegnamento delle lingue e delle letterature (Attributing sense to verbal texts. Aspects of languages and literatures teaching)* (Petőfi [2004] 2012).

with Ugo Volli, Carla Mareello, Manuel Barbera and some PhD students² of the above-mentioned Doctoral School.

Mareello, introducing the round table, observed that Text Linguistics and Semiotics have always had good relationship, stronger thanks also to the work of scholars like J. S. Petőfi (see Petőfi [1976] 1977). Text Linguistics and Corpus Linguistics, on the other hand, had, in 2004, looser connections, which appeared tighter only to few researchers.

1. How Matters Stood

Corpus Linguistics was born to study collocations at short distance. This was emphasized by Doug Biber among others; this same author, however, already in the Nineties lamented that «few studies have exploited the corpora to analyze characteristics of texts rather than characteristic of sentences» (Biber and Finegan 1991: 209) and it was so in spite of the fact that the first corpora were collections of rather long written texts.

Corpus Linguistics began as a branch of Computational Linguistics and was interested in word combination in a sentence range, though investigations at the beginning were not supported by corpora large enough to be representative. Since corpus linguistics made its first steps in the externalist frame, and since the years were what they were, the connection with behaviourism naturally arose.³ Therefore, the “Great Generative War” began; as Leech wrote «Chomsky had, effectively, put to fight the corpus linguistics of the earlier generation» (Leech 1991: 8). McEnery and Wilson (1996) in their well-known handbook still presented corpus linguistics as a radical form of antigerativism.

In his book *Transformationsgrammatiken und eine ko-textuelle Texttheorie* (1971) Petőfi tried to trace a synthesis of the classical transformational theory and of “other” transformational theories such as Fillmore’s case grammar, generative semantics⁴ or Žolkovskij and Mel’čuk’s “Semantic synthesis”.⁵

² Adriano Allora, Stefania Ferraris, Alessandro Panunzi were at that time PhD students of the PhD School in Linguistics, Applied Linguistics and Linguistic Engineering at the University of Turin.

³ «When did modern corpus linguistics begin? Should we trace it back to the era of post-Bloomfieldian structural linguistics in the USA? This was when linguists (such as Harris and Hill in the 1950s) were under the influence of a positivist and behaviourist view of the science and regarded the ‘corpus’ as the primary explicandum of linguistics» (Leech 1991: 8).

⁴ Linguists mentioned by Petőfi under the label of generative semantics are P. Postal, E. Bach and J. McCawley.

He faced the problem of overcoming the sentential domain (*Überschreiten der Satzdomäne*) and the problem of semantic implications. In that first stage, he planned a grammatical theory of text, which was soon going to develop from co-textual into con-textual theory, but corpora were not considered in his theory a viable witness. As he pointed out clearly (see Petőfi 1974), for him TeSWeST (acronym for *TextStruktur-WeltStruktur-Theorie*) had to be *corpus specific* in a very different way from what we understand when we speak of corpus specificity today. For him *corpus* was to be meant as an *infinite* set of texts sharing peculiar structural features. He added that to analyse a *finite* corpus of texts was useful only in order to develop TeSWeST so that it was flexible and comprehensive enough to describe different fields. He was more interested in the semantic-extensional component of his theory and not in the purely linguistic form of texts.

Observing from outside—so to say—models proposed by text linguists up to that date, Levinson (1983) maintained that they were inductive and based on introspection, so putting textualists in broadly the same field as generativists. Therefore, with corpus linguists in the antigenerativist field and text linguists in the generativist ranks, it was too early to openly deal with text corpus linguistics. Moreover, in the Eighties corpus-driven linguistics à la Sinclair was not really interested in text features and it appeared to lay itself open to criticism from more than one linguistic front.

After decades of textual studies progressively more corpus-based, and above all based on more accurately collected corpora, we thought that it was time to dare to speak of *text corpus linguistics*. During the Torino round table Petőfi, faithful to his ideas, showed very little enthusiasm for corpus linguistics, saying that research based on corpora is not totally reliable, or in any case no more reliable than the results linguists can reach via introspection with much less work.⁶

During the round table, our PhD students and ourselves reported our toils with the construction of corpora from newsgroups in various languages and proudly produced first results of sample studies based on

⁵ Žolkovskij and Mel'čuk (1967) was then known to the few who could read Russian. It was one of the first western (i.e. outside the academic milieu then dominated by Soviet Union) mentions of the work that Mel'čuk afterwards developed into Meaning-Text Theory; Mel'čuk and Žolkovskij (1984) is the fully fledged dictionary in which the two linguists published the results of their 1967 reflections.

⁶ The round table was recorded and in this paper we are quoting from our transcription of Petőfi's intervention published in June 2018 in *RiCognizioni*, an online journal.

NUNC corpora.⁷ Petőfi, with the great sensibility he always showed towards younger researchers, warned us against the danger of having PhD students starting from scratch instead of building upon what was already collected.⁸

Texts for corpora need to be very carefully selected and Petőfi said that «[he] would prefer if a video recording was added to corpus analysis, as you cannot always prove the existence of irony just by relying on transcripts» (Petőfi [2004] 2018, our translation).

The mention of *videotapes* did not surprise us because we knew Petőfi's interest for visual texts. However, when we happened to read an interview released by Chomsky the same year, we wondered whether we had decided not to pay due attention to Petőfi's words.

Chomsky confirmed his negative opinion about the scientific value of corpus linguistics in this way: «suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. [...] So if results come from study of massive data, rather like videotaping what's happening outside the window, fine—look at the results. I don't pay much attention to it. I don't see much in the way of results» (Chomsky, in Andor 2004: 97).

Petőfi's argument was more structured than Chomsky's, nonetheless, linking his reflections to our interests for lexical statistics, he reminded us of the witticism by which his Hungarian statistics professor used to open his course: «statistics is a right set of inaccurate data» (Petőfi [2004] 2018, our translation).

2. Text, Context and Co-text

If we just remove an unnecessary anti-behaviourist flavour, the problem declares itself for what it is: a matter of context.

⁷ NUNC corpora are freely accessible at <http://www.bmanuel.org/projects/ng-HOME.html> (and <http://www.corpora.unito.it>). For a printed version of those first results, see Barbera, Corino and Onesti (2007b).

⁸ Petőfi had been dealing with such topics for thirty-five years and he realized that every five-six-seven years researchers start again from scratch, sometimes because it is necessary to label a research differently in order to be funded. As for corpora analysis Petőfi observed that German researchers did huge work and then abandoned it because they had not recorded all the data necessary to conduct a sound analysis.

Due to copyright problems, corpora with fully available texts were few and very expensive. It was mainly a practical matter, but with heavy side effects: unable to reach the context, textualists avoided using corpora. But since a possible solution of this legal conundrum was provided (e.g. Ciurcina and Ricolfi 2007), this tendency was reversed (cf. Barbera 2013b: 20, 67). Things are going better and better thanks to improvements in the metadata rendition, making pieces of co-text available.

Petófi's concern, however, went beyond, and reached the delicate question of the representation of co-text in corpora. According to him, if you want to combine corpora with textual analysis, information about the co-text is necessary, i.e. the verbal context of the sentences or the textual portions to be analysed. Of course, we should not limit ourselves to the co-text. As for the *lexicon encyclopedia*, analysis does not end with the encyclopedia or syntactic analysis. In this case Petófi would work on both inductive and abductive thinking: «when I have seen this and that, the situation looked like that, now I see this other constellation and I think that also in this case there is the same situation» (Petófi [2004] 2018, our translation). Such a position is consistent with his long lasting focus on coherence (cf. Andor and Hölker, this volume).

3. Textology and Multimedia Corpus Linguistics

Petófi remarked that closed hypertexts are different from open ones. What is more, closed hypertexts are made of written texts. In dealing with these hypertexts, the combined use of different media must be taken into account. Petófi thought that there are no monomedial texts, as even a normal text is composed of a conceptual medium and a visual/auditory medium (Petófi [2004] 2018). It is true that in the linguistic studies of the third quarter of 20th Century only considered “normal” written (printed) texts,⁹ but it was, and is, obviously insufficient. Now we do know that texts have many more aspects than just the linguistic one and Petófi, as always, was not only a backwoodsman (as Peirce would have said it) in this field. With his Semiotic Textology, he reached what is, by now, the best theorization available,¹⁰ the only theory which can holistically face

⁹ It happened even with the more theoretically-minded linguists: Roy Harris, in his renowned 1984 paper, dealt with “unconventional texts” such as wall writings, product labels; all written texts which could not be fully understood without a description of their physical support and its surrounding.

¹⁰ In the words of Giuffrè and Scibetta (2014), it is «the most complete theory». For some presentation of the theory see at least Petófi (2004) and the earlier Petófi and La Matina (1994) and Petófi and Vitacolonna (1996); a very useful reference

communication and interpretation, combining the grammatical, semantic and pragmatic components in one semiotic dimension.

Petőfi's textology avails itself of formal and model-theoretic (Montague) logics, extensional and intensional semantics, structural and procedural analysis and the outcome of the Speech Acts Theory. It is a rigorously planned combination of linguistics and semiotics *stricto sensu*, epistemology, psychology, anthropology, etc. The texts are viewed both in their internal formal structure and in their external (but not peripheral) relations with the actors of the communicative act. Semiotic Textology seems very near to the best fulfilment of Peirce's program, even though Peirce's works are mentioned neither in Petőfi first works nor later.

Petőfi's textology was established before the world-wide success of Corpus Linguistics; the distances reduced since when Petőfi commented unfavourably on it; but still are there. Serious textological concerns, indeed, were well taken in account in what is, in our opinion, the most carefully worded (and documented) definition of corpus available up to now:

Collection of texts (written, oral or multimedia) or parts of texts in finite number, in digital format, homogeneously processed (i.e. tokenized and equipped with adequate mark-up) so to be computationally managed and queried. If the corpus, as often happens, aims to be a tool for natural languages description, the texts collected are to be authentic and representative.¹¹ (Barbera, Corino and Onesti 2007a: 70, our translation)

Currently, however, the study and the construction of multimedia corpora (or *multimodal corpora*¹² as are technically called) are «still in [their] relative infancy» (Knight 2011: 392).

In the 2004 Turin round table, Marengo observed that Semiotics of multimedia text and Corpus Linguistics did not seem to have great

is Borreguero Zuloaga (2005), lamentably unpublished but circulating among Petőfian followers.

¹¹ «Raccolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi» (Barbera, Corino and Onesti 2007a: 70).

¹² Following Dawn Knight, Abuczki Ágnes and Esfandiari Baiat Ghazaleh (Knight 2011: 392 and Abuczki and Ghazaleh 2013: 88) we can define with Foster and Oberlander (2007: 307-308), *multimodal corpus* as «an annotated collection of coordinated content on communication channels including speech, gaze, hand gesture and body language, and is generally based on recorded human behaviour».

connections. In fact, increasing such connections should have been planned: multimedia corpora were just beginning to appear [it was fourteen years ago]. However, it is not only their construction process that needs to be studied, also (at least for the Italian language) they seem to be lacking certain features.¹³ Most of them do not have a spoken text aligned with a visual text, there are only transcription of broadcastings.¹⁴ It may be that the flourishing production of commercial and amateur videos equipped with soundtrack will provide that variety of visual texts required by Semiotic Textology; the evolution of voice recognition software will allow easy access to the linguistic content of such visual texts, dispensing us from transcriptions (or providing semi-automatic transcriptions).

The European scene as well does not appear to be well versed in this field, but the topic is hot in the corpus linguistics community.¹⁵ The most noteworthy approximation¹⁶ to these challenges in our opinion is the neat Lancaster Corpus of Children's Project Writing (LCCPW)¹⁷, a true pioneer since it was launched in 1996 (and is constantly adjourned). Indeed, real multimodal corpora, as we said, are still very young, and the worldwide review offered by Dawn Knight is suggestive but very meagre; some of her insights and suggestions, however, run in the direction that Petöfi would have advocated.

Marellò in the 2004 Turin round table identified two main needs: (1) a multimodal query language, able to start with iconic elements and aiming to linguistic ones, and also a multimodal query language able to search mingled elements; (2) a mark-up able to account for interwoven coding and open to a new kind of texts, also only machine-readable texts. In 2018 we have all the above mentioned. For the latter point, Marellò focused on the opportunity not to leave solely IT specialists with the burden to decide ontology and taxonomies to be used for semiotic mark-up. On the contrary, she believed, and still maintains, that it was, and still is,

¹³ Properly speaking we have to deal with what in Italian are called *corpora di trasmesso* (broadcast language corpora), namely a particular kind of oral corpora.

¹⁴ The more significant example (at least known to Barbera [2013c] 2015a: 124) is the LIT (*Lessico di Italiano Televisivo* | Italian Broadcasted Lexicon) directed by Nicoletta Maraschio (whose homepage is:

http://www.italianotelevisivo.org/contenuti/36/banche_dati); LIT however is planned to offer on line spoken/transcribed texts aligned with their visual recordings.

¹⁵ A workshop on *Multimodal Corpora: Computer Vision and Language Processing* was held in LREC Congress (May 2016, Portoroz, Slovenia).

¹⁶ There aren't videotapes: only images linked to written texts.

¹⁷ Homepage: <http://www.lancaster.ac.uk/fass/projects/lever/>.

also the text linguist's duty to establish criteria for the treatment of metadata.

Notwithstanding the progress made, Petőfi's remark that multimedial text semiotics and corpus linguistics do not have the relations they should have holds true: why is it so?

4. Corpus Linguistics and Text Linguistics

Multimedial corpora can provide a fuller answer, but they are still more a promise than a reality. However, something in the right direction also outside multimedia corpora has been achieved.

Nowadays even the choice of the morphosyntactic tagset is linked to the text that has to be tagged, and intelligent browsers use stochastic techniques, which include textual features (where exactly that word is, with which words, at what distance). What is more, websemantics¹⁸ is now the modern web-oriented embodiment of what *Textsorten / Texttypologie* used to be, and it is also a partial answer to the encyclopaedia problem posed by Petőfi. Therefore, the union between corpus linguistics and text linguistic is strengthened, sure and undeniable at least *de facto*, even though most "corpus-based researchers" have no clear perception of their textual engagement at a theoretical level. This is even truer for those who think to conduct a corpus-driven research. To them we should repeat what Petőfi wrote in the Sixties and maintained during the following fifty years: *textual linguistics* is an inappropriate expression inasmuch linguistics alone cannot account for *all* the elements relevant in a text.

Ferrari and Manzotti in 2002 in their survey for Società di linguistica italiana asked «Where is Italian text linguistics going?». After three lustra, we can see that text linguistics concepts have entered contemporary general linguistics handbooks and have permeated all the research fields and linguistic practices (L1 and L2 Italian teaching, academic writing teaching, translation studies, sociolinguistic analysis, typology, etc.). There are deep analyses of single textual phenomena, and educational uses of such analyses that have few equals for other languages, we dare say. However, there was also a lack in more theoretical and general research aimed at defining boundaries and goals of Text Linguistics.

This situation has recently changed since the appearance in 2014 of Angela Ferrari's handbook; as she points out it is not *text linguistics*, but

¹⁸ Websemantics and the like are the natural evolution of the meaning-in-context problem raised by Textology. For, instead, its substantiation as cognitive frames cf. Pelyvás 2013 and the bibliography here reported.

linguistics for the text,¹⁹ well aware that it contributes to text studies together with other branches of learning. In the light of the shifted focus proposed by Angela Ferrari, it might be wise to speak of corpus-based linguistics for text theory, rather than corpus-based text linguistics as our title suggests. Even more when our studies are dealing with (almost monomedial) written text corpora.

Instances of how Petőfi dealt with literary texts in the frame of semiotic textology can be found in his book *Scrittura e interpretazione. Introduzione alla testologia semiotica* (2004), where he gathered and reworked some detailed analyses of different types of texts: poems, pieces of prose by Borges and Calvino. As Borreguero (2011: 66) noted, when Petőfi was asked why he began to deal with linguistics, he answered that his interest was focused on literary communication as the most complex manifestation of verbal communication (Petőfi 1991: 7). Actually, Petőfi's analyses are the clearest demonstration of what can be done with and beyond linguistics for the text.

In Marellò (2011) we can find an attempt to interpret texts of multimodal VALICO corpus.²⁰ This is not from the linguistic point of view of acquisitional varieties of L2 Italian, but from the point of view of coherence, in order to investigate the relationship between what was designed in the pictures but not mentioned in the corresponding elicited story. Or, conversely, in order to investigate what was not in the picture but was expressed in the text and probably belonged to the interpretative "luggage" of the authors of VALICO texts. Using Petőfian terminology, the internal text coherence of the fragment of the world created by the text and the relationship between two costringencies,²¹ that of the iconic *vehiculum* and that of the written linguistic *vehiculum*. The fact of having hundreds of texts to compare among them (knowing that they were elicited by the same iconic stimulus) allows us to identify hyperonymic texts and interpretations of first and second (=symbolic) grade.

¹⁹ In her own words «la linguistica del testo [...] è in definitiva una linguistica *per il testo*» (Ferrari 2014: 15).

²⁰ Freely accessible at <http://www.valico.org>, VALICO is a learner's corpus made of short texts written by students of Italian as L2. It is multimodal because texts were written starting from an iconic input, a comic strip without words.

²¹ Costringency is the continuous and complete network of state of affairs which constitute a fragment of world (Petőfi 2004: 100)

5. Corpus Linguistics for Text Theory

The main problems of the troubled relations between Corpus Linguistics and Text Linguistics can perhaps be reduced to the general difficulties tied to implementing and querying multimedial corpora: Petőfi, as always, was simply ahead of his times.

The relations of corpus linguistics and generative grammar, even more troubled in the past, meanwhile are more relaxed.

For one side, the polemics have naturally dampened, since linguists all over the world have found the use of corpora very useful, notwithstanding Chomsky's firm opinions. Corpora usefulness for non-linguists is even more obvious, and Chomsky himself (the politically engaged man, not the linguist, of course! See Chomsky and Herman 2002) resigned to using them.²² It is not by chance that in McEnery and Hardie (2012), the handbook's natural successor of McEnery and Wilson ([1996] 2001), the polemical setting has been conveniently muted.

The arguments of the dispute have deflated while transferring themselves in a battlefield where behaviourism was only an imported good, and almost never a very popular one: viz. the Italian tradition, which always was a "corpus based" one, as Sabatini (2006) recognized.

Barbera has advocated this move several times, more pointedly in Barbera ([2012] 2013), also developing the link with the older tradition of philological linguistics (cf. Barbera 2009: 23). He showed that formulating matters in Saussurian terms helps a lot since, from a linguistic point of view, a corpus is a collection of speech acts (*parole*), and by using them, one can trace back the related states of the *langue*. It is almost certain that not all the elements of a *langue* are displayed in a corpus: it is rather the use, as it is witnessed by corpora, that states the *langue*, even though corpora—being finite by definition—will always be representative of just a part of it (Barbera 2013b: 19). If we do not see corpora as haphazard collections of performances but as natural collections of *actes de parole*, statistical results might be a foundation of state of *langue*.

Chomsky (rather venomously) said that «any natural corpus will be skewed. Some sentences will not occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list» (Chomsky [1958] 1962: 159). But it is not matter of unwanted and unrecoverable "skewness": a corpus unavoidably has

²² To back up his research about the biased use of *genocide* in the five major printed media in USA. Mainstream media used in the tabulation were Los Angeles Times, the New York Times, Washington Post, Newsweek, Times.

limitations which are connatural to any natural text. It cannot be but incomplete, without being a “list”. And when Corpus Linguistics is made with well-balanced corpora of true texts, its incompleteness is part of the game itself, not an unwanted side-effect.

Bibliography

- Abuczki, Ágnes & Ghazaleh, Esfandiari Baiat (2013). An Overview of Multimodal Corpora, Annotation Tools and Schemes. *Argumentum IX*, pp. 86-98.
- Andor, József (2004). The Master and his Performance: An Interview with Noam Chomsky. *Intercultural Pragmatics I*, pp. 93-111.
- Barbera, Manuel (2009). *Schema e storia del “Corpus Taurinense”*. *Linguistica dei corpora dell’italiano antico*. Alessandria: Edizioni dell’Orso.
- ([2012] 2013). Per una soluzione teorica e storica dei rapporti tra grammatica generativa e linguistica dei corpora. *7es Journées suisses de Linguistique. L’empirie en linguistique: variété et complexité des approches*. Lugano, Università della Svizzera italiana, 13-14 settembre 2012, Société Suisse de Linguistique = “Schweizerische Sprachwissenschaftliche Gesellschaft / Société Suisse de Linguistique (SSG/SSL) – Archive”, <<http://www.sagw.ch/fr/ssg/taetigkeiten/7e-Giornate-svizzere-della-Linguistica.html>>. Republished in M. Barbera, *Molti occhi sono meglio di uno: saggi di linguistica generale 2008-12*, [Milano], Qu.A.S.A.R., pp. 27-45.
- (2013a). *Molti occhi sono meglio di uno: saggi di linguistica generale 2008-12*, [Milano], Qu.A.S.A.R.
- (2013b). *Linguistica dei corpora e linguistica dei corpora italiana. Un’introduzione*. Milano: Qu.A.S.A.R.. Ebook disponibile a <<http://www.bmanuel.org/man/cl-HOME.htm>>.
- ([2013c] 2015a). La linguistica dei corpora in Italia all’alba del terzo millennio (pp. 581-598). In Gabriele Iannaccaro (ed.), *La linguistica italiana all’alba del terzo millennio (1997-2010)*. Roma: Bulzoni, “Pubblicazioni della Società di linguistica italiana [SLI]” 58, vol. 2. . Updated version in M. Barbera, *Quanto più la relazione è bella. Saggi di storia della lingua italiana 1999-2014*, Torino - Tricase (LE), bmanuel.org - Youcanprint Self-Publishing, “bmanuel.org glottologica et philologica” series maior 1, pp. 117-138.
- (2015b). *Quanto più la relazione è bella. Saggi di storia della lingua italiana 1999-2014*, Torino - Tricase (LE), bmanuel.org - Youcanprint

- Self-Publishing, “bmanuel.org glottologica et philologica” series maior 1.
- Barbera, Manuel, Corino, Elisa & Onesti, Cristina (2007a). Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup (pp. 5-88). In Manuel Barbera, Elisa Corino & Cristina Onesti (eds.), *Corpora e linguistica in rete*. Perugia, Guerra Edizioni, “L’officina della lingua. Strumenti” 1. .
- Barbera, Manuel, Corino, Elisa & Onesti, Cristina (eds.) (2007b). *Corpora e linguistica in rete*. Perugia, Guerra Edizioni, “L’officina della lingua. Strumenti” 1.
- Biber, Douglas & Finegan, Edward (1991). On the Exploitation of Computerized Corpora in variation studies (pp. 204-220). In Karen Aijmer & Bengt Altenberg, (eds.), *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London - New York: Longman.
- Borreguero Zuloaga, Margarita (2005). *De la gramática del texto a la textología semiótica: aproximaciones al proceso de interpretación textual* (unpublished Ph. D. dissertation). Madrid: Universidad Complutense de Madrid.
- . (2011). El texto literario como eje de la reflexión textual. Primeras incursiones de János Sándor Petőfi en la teoría del texto (1961-1969) (pp. 63-88). In Klaus Hölker and Carla Marelllo (eds.), *Dimensionen der Analyse von Texten und Diskursen. Festschrift für János Sándor Petőfi zum achtzigsten Geburtstag | Dimensioni dell’analisi di testi e discorsi. Festschrift per János Sándor Petőfi in occasione del suo ottantesimo compleanno*. Münster: LIT Verlag..
- Chomsky, Noam ([1958] 1962). A Transformational Approach to Syntax. Paper presented at the *3rd Texas Conference on Problems of Linguistic Analysis in English*, then collected in A. Hill (ed.), *Third Texas Conference on Problems of Linguistic Analysis in English: May 9-12, 1958*. Austin: University of Texas, “Studies in American English”. , pp. 124- 158, also with the *Discussion*.
- Chomsky, Noam & Herman, Edward S. (2002). *Manufacturing Consent: The Political Economy of the Mass Media*. New York: Random House USA Inc.
- Ciurcina, Marco & Ricolfi, Marco (2007). Le Creative Commons Public Licences per i corpora. Una suite di modelli per la linguistica dei corpora (pp. 127-132). In Manuel Barbera, Elisa Corino & Crisitna Onesti (eds.), *Corpora e linguistica in rete*. Perugia, Guerra Edizioni, “L’officina della lingua. Strumenti” 1.
- Corino, Elisa, Marelllo, Carla & Onesti, Cristina (eds.) (2006). *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006*

- | *Proceedings of the XII EURALEX International Congress. Torino, Italia, 6th-9th September 2006*, 2 vols. Alessandria: Edizioni dell'Orso.
- Ferrari, Angela (2014). *Linguistica del testo. Principi, fenomeni, strutture*. Roma: Carocci editore, "Manuali universitari" 151.
- Ferrari, Angela & Manzotti, Emilio (2002). Linguistica del testo (pp. 413-454). In Cristina Lavinio (ed.), *La linguistica italiana alle soglie del 2000: 1987-1997 e oltre*. Roma: Bulzoni, "Pubblicazioni della Società linguistica italiana" 44.
- Foster, Mary Ellen & Oberlander, Jon (2007). Corpus-based Generation of Head and Eyebrow Motion for an Embodied Conversational Agent. *Language Resources and Evaluation XLI*: 3/4, pp. 305–323
- Giuffrè, Mauro, (ed.) (2011). *Studies in Semiotic Textology in Honour of János S. Petőfi. Sprachtheorie und germanistische Linguistik*, Supplement 1.
- Giuffrè, Mauro & Scibetta, Salvatore Alessandro (2014). Semiotic Textology as the "Grundlagenwissenschaft" for Philology. János S. Petőfi's scientific inheritance, *Sprachtheorie und germanistische Linguistik XXIV*: 2, pp. 183-210.
- Harris, Roy (1984). The Semiology of Textualization. *Language Sciences VI*, 271-286. Then collected in R. Harris (ed.), *The Foundations of Linguistic Theory. Selected Writings*, edited by Nigel Love, London & New York: Routledge, 1990, pp. 210-226.
- Hölker, Klaus & Marellò, Carla (eds.) (2011). *Dimensionen der Analyse von Texten und Diskursen. Festschrift für János Sándor Petőfi zum achtzigsten Geburtstag | Dimensioni dell'analisi di testi e discorsi. Festschrift per János Sándor Petőfi in occasione del suo ottantesimo compleanno*. Münster: LIT Verlag.
- Knight, Dawn (2011). The Future of Multimodal Corpora. *Revista Brasileira de Linguística Aplicada XI*: 2, pp. 391-415.
- Leech, Geoffrey (1991). The State or the Art in Corpus Linguistics (pp. 8-29). In Karen Aijmer & Bengt Altenberg (eds.), *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London - New York: Longman.
- Levinson, Stephen C. (1983). *Pragmatics*. Cambridge: Cambridge University Press, "Cambridge Textbooks in Linguistics". Italian translation by Marcella Bertuccelli Papi, *La pragmatica*, Bologna, Il Mulino, 1985 "Strumenti".
- Marellò, Carla (2011). Interpretare testi scritti composti a partire da storie diseguate (pp. 283-304). In Klaus Hölker & Carla Marellò (eds.), *Dimensionen der Analyse von Texten und Diskursen. Festschrift für János Sándor Petőfi zum achtzigsten Geburtstag | Dimensioni*

- dell'analisi di testi e discorsi. Festschrift per János Sándor Petőfi in occasione del suo ottantesimo compleanno.* Münster: LIT Verlag.
- McEnery, Tony & Hardie, Andrew (2012). *Corpus Linguistics. Method, Theory and Practice.* Cambridge - New York etc.: Cambridge University Press, "Cambridge Textbooks in Linguistics".
- McEnery, Tony & Wilson Andrew ([1996] 2001). *Corpus Linguistics. An Introduction.* Edinburgh: Edinburgh University Press, "Edinburgh Textbooks in Empirical Linguistics".
- Péter, Pelyvás (2013). Meaning at the Level of Discourse: from Lexical Networks to Conceptual Frames and Scenarios. *Officina textologica* 18, pp. 14-39.
- Petőfi, János S. (1971). *Transformationsgrammatiken und eine ko-textuelle Texttheorie. Grundfragen und Konzeptionen.* Frankfurt am Main: Athenäum Verlag.
- ([1976] 1977). *Some Remarks on the Grammatical Component of an Integrated Semiotic Theory of Texts.* Bielefeld University, unpublished paper. Italian translation by Carla Marelllo, Osservazioni sul componente grammaticale d'una teoria semiotica integrata dei testi (pp. 224-247). In M. E. Conte, ed., *La linguistica testuale.* Milano: Feltrinelli Economica, "SC/10 Readings" 4.
- (1991). *A humán kommunikáció szemiotikai elmélete felé (Szövegyelvészet, Szemiotikai textológia) | Towards a Semiotic Theory of the Human Communication (Text Linguistics, Semiotic Textology).* Szeged: Gold Press.
- Petőfi [*sic*], János S. ([1988] 1996). La lingua come mezzo di comunicazione scritta: il testo. Urbino, Centro internazionale di semiotica e linguistica dell'Università di Urbino. Reprinted in János S. Petőfi & Luciano Vitacolonna (eds.), *Sistemi segnici e loro uso nella comunicazione umana. 3. La testologia semiotica e la comunicazione multimediale.* Macerata: Università di Macerata, "Dipartimento di filosofia e scienze umane. Quaderni di ricerca e didattica" 17, 1996, pp. 66-107. English version in *An Encyclopedia of Language*, edited by N[eville] E. Collinge, London - New York, Routledge, 1990, ¶ 7, pp. 207-243.
- Petőfi, János S. ([2004] 2012). Il conferimento del senso ai testi verbali. Aspetti dell'insegnamento delle lingue e delle letterature, *Lectio magistralis* held in Turin (February 27th, 2004). Published in Elisabetta Soletti & Cristina Onesti (eds.), *Pensieri e parole del Novecento.* Alessandria: Edizioni dell'Orso, 2012, pp. 167-181.

- Petőfi [sic], János S. (2004). *Scrittura e Interpretazione. Introduzione alla Testologia Semiotica dei testi verbali*. Roma: Carocci, “Università”.
- Petőfi, János S. ([2004] 2018). Talk at the Tavola Rotonda. *Linguistica del testo, semiotica del testo, linguistica dei corpora: quali rapporti?* Turin, 27 February 2004. Published in *RiCognizioni V*: 9 (2018).
<<http://www.ojs.unito.it/index.php/ricognizioni>>
- Petőfi, János S., Békési, Imre & Vass, László (eds.) (1994). *Szemiotikai Szövegtan. 7. A multimediális kommunikátumok szemiotikai textológia megközelítéséhez*. Széged, JGYTF Kiadó, “Acta Academiae Pedagogicae Szegediensis (A Szegedi Pedagógiai Főiskola évkönyve). Series linguistica, litteraria et aethetica”.
- Petőfi, János S. & La Matina, Marcello (1994). Egy általános szemiotikai textológia centrális aspektusai (pp. 97-128). In János S. Petőfi, Imre Békési & László Vass (eds.), *Szemiotikai Szövegtan. 7. A multimediális kommunikátumok szemiotikai textológia megközelítéséhez*. Széged, JGYTF Kiadó, “Acta Academiae Pedagogicae Szegediensis (A Szegedi Pedagógiai Főiskola évkönyve). Series linguistica, litteraria et aethetica”.
- Petőfi [sic], János S. & Vitacolonna, Luciano (eds.) (1996). *Sistemi segnici e loro uso nella comunicazione umana. 3. La testologia semiotica e la comunicazione multimediale*. Macerata: Università di Macerata, “Dipartimento di filosofia e scienze umane. Quaderni di ricerca e didattica” 17.
- Sabatini, Francesco (2006). La storia dell’italiano nella prospettiva della corpus linguistics (pp. 31-37). In Manuel Barbera, Elisa Corino & Crstina Onesti (eds.), *Corpora e linguistica in rete*. Perugia, Guerra Edizioni, “L’officina della lingua. Strumenti” 1. .
- (2007). Storia della lingua italiana e grandi corpora. Un capitolo di storia della linguistica (pp. xij-xvj).. In Manuel Barbera, Elisa Corino & Crstina Onesti (eds.), *Corpora e linguistica in rete*. Perugia, Guerra Edizioni, “L’officina della lingua. Strumenti” 1
- Žolkovskij, Alexander K. & Mel’čuk, Igor A. (1967). A[лександр] K[онстантинович] Жолковский - И[горь] A[лександрович] Мельчук, О семантическом синтезе | I[gor'] A[lexandrovič] Mel’čuk - A[lexander] K[onstantinovič] Žolkovskij, O semantičeskom sinteze (On Semantic Synthesis). *Problemy kibernetiki* 19, pp. 177–238.
- Žolkovskij, Alexander K. & Mel’čuk, Igor A. (1984). И[горь] A[лександрович] Мельчук - A[лександр] K[онстантинович] Жолковский, Толково-комбинаторный словарь современного русского языка. Опыты семантико-синтаксического описания русской лексики | A[lexander] K[onstantinovič] Žolkovskij, I[gor']

A[lexandrovič] Mel'čuk, Tolkovo-kombinatornyj slovar' russkogo jazyka (Explanatory and Combinatorial Dictionary of Modern Russian). *Wiener Slawistischer Almanach*, Sonderband 14, p. 992.