Introduction to the special issue on annotated corpora

Marie Candito, Mark Liberman

Introduction to the special issue on annotated corpora

Marie Candito* — **Mark Liberman****

* *LLF - Université Paris Diderot / CNRS*
** *University of Pennsylvania*

ABSTRACT. *Annotated corpora are increasingly important for linguistic scholarship, science and technology. This special issue briefly surveys the development of the field and points to challenges within the current framework of annotation using analytical categories as well as challenges to the framework itself. It presents three articles, one concerning the evaluation of the quality of annotation, and two concerning French treebanks, one dealing with the oldest project for French, the French Treebank, the second concerning the conversion of French corpora into the cross-lingual framework of Universal Dependencies, thus offering an illustration of the history of treebank development worldwide.*

RÉSUMÉ. *Les corpus annotés sont toujours plus cruciaux, aussi bien pour la recherche scientifique en linguistique que le traitement automatique des langues. Ce numéro spécial passe brièvement en revue l'évolution du domaine et souligne les défis à relever en restant dans le cadre actuel d'annotations utilisant des catégories analytiques, ainsi que ceux remettant en question le cadre lui-même. Il présente trois articles, l'un concernant l'évaluation de la qualité d'annotation, et deux concernant des corpus arborés du français, l'un traitant du plus ancien projet de corpus arboré du français, le French Treebank, le second concernant la conversion de corpus français dans le schéma interlingue des Universal Dependencies, offrant ainsi une illustration de l'histoire du développement des corpus arborés.*

KEYWORDS: *Annotated corpora, Resources for NLP, Linguistic resources*

MOTS-CLÉS : *Annotation de corpus, Ressources pour le TAL, Ressources linguistiques*

TAL. Volume 60 – n°2/2019, pages 7 à 17

1. From corpus-based theology, anthropology, and lexicography to modern NLP

Long before the invention of digital computers, the collection and annotation of corpora began as efforts to support the interpretation of culturally important texts, and continued as efforts to document language use. We'll introduce this special issue with a brief survey of some of this history, before providing some historical notes on annotated corpora for NLP in section 1.1, and for research in linguistics in section 1.2.

One well-known example of text augmented with interpretation comments is the Talmud, which, as Mielziner (1903) explains, "consists of two distinct works, the *Mishna*, as the text, and the *Gemara* as a voluminous collection of commentaries and discussions of that text". A different sort of religiously-motivated example is Strong (1890), which assigns a number to each of 8,674 Hebrew "root words" (what we would call "lemmas") used in the Old Testament, and similarly to each of 5,624 Greek lemmas in the New Testament, and annotates each relevant word of the English King James Version with the number of the source-language word. This allows someone to read the original texts, in some sense, even if they have little or no knowledge of the source languages. And because the whole thing is indexed as a concordance, they can find and compare all of the passages that use a given English, Hebrew, or Greek word.

An approach more similar in design to contemporary corpus annotation is the tradition of interlinear text, represented in a fully mature form in Müller *et al.* (1864). As that work's title states, it presents three layers of annotation for each word of the Sanskrit text of the *Hitopadeśa*: "transliteration, grammatical analysis, and English translation". As with the Talmud and Strong's concordance, Müller's work was based on a text that had existed for hundreds of years. In the late 19th and early 20th century, interlinear annotations of newly-collected texts became a standard tool in the development of linguistic anthropology and linguistic documentation. Franz Boas and his many followers saw the collection and annotation of texts as central to the understanding of languages and cultures (Boas and Hunt, 1905; Epps *et al.*, 2017).

Roberto Busa's 1946 dissertation (Busa, 1949) dealt with the concept of "presence" in the thought of Thomas Aquinas. Busa came to believe that he needed to understand the shades of meaning associated with Aquinas's use of the Latin preposition *in*, and based his dissertation on 10,000 examples of this usage, collected and written by hand on file cards. As he completed that work, Busa began to dream of a set of machine-readable cards a thousand times larger, from which automatic sorting could create unlimited opportunities for such corpus-based theology. In 1949, even before the first primitive digital computers became generally available, Busa persuaded IBM to support the creation of the *Index Thomisticus*, a complete lemmatized concordance to everything Aquinas wrote (Jones, 2016; Winter, 1999). The millions of punched cards constituting this corpus were completed in 1967 (Busa, 1974).

A similar transition from file cards to punched cards (and digital tape and onwards) took place in the field of lexicography. Starting in 1879, James Murray's "reading programme" turned mountains of books into thousands of "slips" documenting word usage, eventually organized into the *Oxford English Dictionary* (Winchester, 1998).

A similar program was carried out by the Merriam-Webster company in creating their series of dictionaries. In the early 1960s, Kučera and Francis created the Brown corpus (Kučera and Francis, 1967), a million-word balanced corpus of written American English meant to support "computer-based research in the English language"; and just a few years later, Houghton-Mifflin used the Brown corpus as the citation base for the first edition of the *American Heritage Dictionary* (Morris, 1969).

The idea of basing dictionaries and other forms of language documentation on large representative digital text corpora was soon adopted widely, since it was essentially a computational enhancement of methods that had long been in use. In Sweden, the pioneering work of Sture Allén at the University of Gothenburg in the 1960s led to Press-65, an electronic text corpus of one million words of newspaper text (Allén, 1968), and the establishment of Språkbanken (the Swedish Language Bank) in 1975. Also in the 1970s, researchers at Lund University created Talbanken, a syntactically-annotated corpus of Swedish (Einarsson, 1976a; Einarsson, 1976b). The Czech Academic Corpus project started in the 1970s at the Institute of the Czech Language [1]. The primary goal of this project was to create a corpus that would contain manual annotation of morphology and syntax of Czech, as a base for building a frequency dictionary. In the UK, the Collins Birmingham University International Language Database (COBUILD) project, begun in 1980 and funded by the Collins publishing company, formed the basis for the *Collins COBUILD Dictionary*, first published in 1987, and for a series of other reference works (Sinclair, 1987; Moon, 2009). And, beginning in 1991, Oxford University Press led a consortium of dictionary publishers, academic research centers, and government organizations in creating the British National Corpus (BNC), an open corpus of 100 million words from sources meant to be representative of spoken and written English in Great Britain in the last decade of the 20th century (Aston and Burnard, 1998). Part-of-speech and sense tagging were involved in the BNC project from the beginning (Atkins, 1993; Leech *et al.*, 1994; Kilgarriff, 1998).

In France, the *Trésor de la Langue Française* ("French language treasury", hereafter "TLF") project started in the 1960s to study French vocabulary usage, at the INALF laboratory in Nancy. As part of this project, the Frantext textual database of literary and technical texts was set up in order to provide examples for the TLF dictionary. After dictionary completion (the 16 volumes were released between 1971 and 1994, and are now an online resource, the "TLFi"), access was given to the textual database, first within the lab via a search engine in 1985, and later via online access in 1998 (according to Montémont [2008]). The current version contains over 5,000 automatically-parsed texts, totaling 250 million words [2]. And the MULTEXT project, funded by the European Commission in the early 1990s (Ide and Véronis, 1994), aimed to provide "generally usable software tools to manipulate and analyse text corpora and to create multilingual text corpora with structural and linguistic markup."

1. For a history of the project, see `http://ufal.mff.cuni.cz/rest/CAC/doc-cac10/cac-guide/eng/html/chapter2.html`.
2. `https://www.frantext.fr/`.

The MULTEXT plan included standards for encoding linguistic annotation, including morphology, syntax, parallel text alignment, and prosody.

1.1. *Corpora for Human Language Technology*

Turning to the use of corpora for human language technology, research on the statistics of messages, by engineers tasked with sending them efficiently turned to have a major impact – perhaps the most important one for the modern use of corpora in NLP, since it has resulted in a flowering of diverse approaches to the annotation of large linguistic datasets, used to train computer systems for tasks far beyond the problems of message encoding and transmission.

Shannon (1948) laid out a theory of optimal transmission over noisy channels, and showed that a simple empirical model could provide an arbitrarily close approximation to the relevant statistics of message sequences. Baum *et al.* (1970) and others provided computationally-efficient extensions of this model to cases where we can observe only a stochastic function of the hypothesized underlying text – and, over the course of the next few decades, such methods were applied to problems of speech recognition (Baker, 1975), machine translation (Brown *et al.*, 1990), part-of-speech tagging (Church, 1989), parsing (Lari and Young, 1990), and many other forms of speech and language analysis.

Starting in the mid-1980s, the US Defense Advanced Research Projects Agency (DARPA) began promoting research in this area, using what has been called the "Common Task method" (Liberman and Wayne, 2020) or "Shared Task method". This research management technique begins with a shared training dataset and an automated quantitative performance metric, with periodic competitive evaluations on test data withheld for the purpose. Because the datasets and evaluation software were published, and because the method succeeded in fostering gradual improvements in the targeted technologies, the Common Task approach has been widely adopted.

And large annotated corpora have been at the center of the process, since the dominant paradigm has been supervised machine learning, in which a system is trained and tested on a body of text or speech in which the desired analysis is explicitly presented (Liberman, 1991). Such analyses include morphosyntactic categories and relations in text, word senses, references to semantically-defined entities, textual evidence for adding information to a "knowledge graph", and many other things. The effectiveness of supervised models crucially depends on the availability of a large volume of annotated corpora. We focus below on the early days of two famous types of annotation, POS-annotated corpora and syntactic treebanks, whose development has been parallel to statistical taggers and parsers.

Perhaps the earliest example of digital corpus annotation was the part-of-speech (POS) tagging of the previously-mentioned Brown Corpus, which was motivated by language documentation and lexicography, and obtained by manually correcting the output of hand-written rules (Greene and Rubin, 1971). POS taggers based on statisti-

cal machine learning (Garside *et al.*, 1987; Church, 1989) were developed in the 1980s and early 1990s, as were broad-coverage statistically-trained parsers (Magerman and Marcus, 1990; De Marcken, 1990). These systems were trained on text collections provided with part-of-speech tags and syntactic analyses – and the same systems were also used in the creation of these treebanks to improve the productivity of human annotators by providing them with an automatically-created draft to correct. Morphosyntactically annotated corpora followed for a wide variety of typologically diverse languages.

The most well known and widely used syntactically-analyzed corpus has been the Penn Treebank (Marcus *et al.*, 1993), whose creation started in the late 1980s. The resulting dataset has been used for thousands of works on English syntax and systems for analyzing it – Google Scholar lists about 19,000 works that reference it. A crucial reason for this popularity (besides the economic importance of the English language) is that the Penn Treebank was made easily available to researchers all over the world, as the Brown corpus had been, in contrast to some other early collections such as the Swedish treebanks, which were tightly held by their creators. Similar treebanks were subsequently built for many typologically diverse languages.

A parallel strand of corpus creation has used syntactic dependencies (Tesnière, 1959) rather than syntactic constituents. From a formal point of view, these two representations are essentially equivalent, but the non-nested syntactic relations that are common in free-word-order languages are more easily represented by (crossing) dependencies than by (discontinuous) constituents. The Prague Dependency Treebank (with a first release in 1998 [Hajič, 1998]) was an important influence on the history of this approach, which has culminated in the Universal Dependencies project [3] (Nivre *et al.*, 2016), an attempt to provide cross-linguistically consistent dependency annotation. This open community effort with over 200 contributors has led to 146 treebanks covering 83 languages (as of version 2.4). Despite the inevitable approximations of the cross-lingual scheme, the resource is abundantly used for typological quantitative research and for cross-lingual parsing.

Over the past decade, so-called "deep learning" methods have achieved better performance than statistical machine-learning methods on most NLP tasks. But these methods are even hungrier for training data, and so the need for large annotated corpora to support Human Language Technology has only increased.

1.2. *Corpora for scientific and scholarly research*

The Child Language Data Exchange System (CHILDES), begun in 1984, established a repository for sharing records of child language acquisition (MacWhinney and Snow, 1985; MacWhinney, 2014), based on a system for discourse notation and coding called CHAT, and including transcripts of caregiver-child interactions expressed

3. `https://universaldependencies.org/`.

in that form, in some cases with audio and/or video recordings. The accumulation of shared material in CHILDES continues, with more than 130 different sources in 26 languages now available. The CHAT format allows for (but does not require) annotation of time codes, pronunciations, dysfluencies and speech errors, prosody, and speech act categories, among other things.

The study of historical syntax has always been corpus based, since historical texts provide the only concrete evidence of how language was used in earlier periods. Over the past twenty years, researchers at universities in the US, the UK, Finland, and Norway have created a series of parsed corpora totaling about 9 million words and covering more than a thousand years of linguistic history (Taylor, 2020). Similar efforts are underway for French, Portuguese, and Spanish.

Since the 1960s, research in quantitative sociolinguistics has been based on statistical modeling of annotation of the speech patterns of people varying in gender, age, location, socio-economic status, and communicative context. In earlier times, recordings and annotations were closely (and informally) held by the researchers that collected them, but, more recently, the culture of the discipline has begun changing in the direction of digital archiving and sharing of research datasets (Kendall, 2008). A similar trend has resulted in the digitization and (partial) publication of the archives of many dialect atlas projects from the past century (Nerbonne, 2009).

Thanks to the influence of psychology, as well as results within theoretical linguistics itself, such as Bresnan (2007), statistical information is increasingly viewed as part of the competence of speakers rather than a mere artefact of linguistic performance. The statistical exploration of corpora thus serves to validate linguistic hypotheses and even to discover new patterns. In such studies, sophisticated linguistic annotation may be needed in order to compile the needed counts.

2. Forward-looking perspective

Looking forward to the future of annotated corpora, we can first comment on some "sociological" aspects, both within the linguistics and NLP research communities. Annotation projects are costly and time-consuming, meaning reusability is crucial. NLP researchers tend to reuse already existing datasets, not only to enable comparison between systems but simply because it is easier. This might explain the emergence of annotation scheme standards, often stemming from English-centered projects, initiated by major American NLP players. Resources in other languages either use the same schemes, even if at a smaller scale, or an original annotation scheme, at the risk of lacking international visibility. Interestingly, the Universal Dependencies project, although initially driven by such players (Stanford, through the Stanford dependencies, and Google, through the universal POS tagset), has fostered a global reflection on an annotation scheme with a multilingual vocation.

From a practical point of view of easing linguistic resource production, there are currently several approaches, in particular using more efficient tools for experts, or us-

ing non-expert annotators. The possibility to leverage a potentially worldwide workforce via crowd-sourcing platforms has modified the economics of resource production. Note though that inequalities between languages are reinforced. In 2014, out of 100 languages, 13 only were considered to have an adequate workforce among the turkers (Pavlick *et al.*, 2014). Ethical concerns are now part of major NLP conferences and the focus of specific workshops or special issues (Fort *et al.*, 2016; Hovy *et al.*, 2017). *Games with a purpose* are an alternative used for varied linguistic resources including annotated corpora, for instance for coreference annotation (Chamberlain *et al.*, 2013) or dependency syntax (Guillaume *et al.*, 2016).

Another research lead is to make better use of the existing resources, in particular with more efficient learning from small datasets. Multilingual learning leverages data in several languages to better model the very same languages or some related ones.

Coping with multilinguality for resource production is indeed a major challenge, both practical and scientific. As already noted, the Universal Dependencies project succeeded in having hundreds of contributors collaborate. The PARSEME project has produced guidelines and data for 20 languages, for verbal multi-word expressions. Finally, annotating multimodal data is another challenge, with the necessity to annotate interconnections between different modalities (speech, gesture, emotional states...).

The framework of annotating corpora using analytic categories (for whatever kind of linguistic concept) is itself challenged, in particular given the current use of continuous representations in neural models. Frontiers between linguistic categories are often difficult to draw sharply (examples of well-known difficulties are the argument/adjunct distinction in syntax, or the adjective versus participle distinction in many languages, including French or German). This inevitably impacts the annotation process. For instance, Plank *et al.* (2014) show that disagreements on POS annotation concern debatable linguistic points rather than random errors, and should be used in the learning phase. These difficulties question the theory, but also the current annotation methodology, in which the resolution of annotation conflicts is the most time-consuming phase.

From the theoretical point of view, this indeterminacy of analytic category frontiers might be justified. Firstly, abandoning rigid frontiers between categories is empirically validated by the success of using continuous representations in neural NLP models, for all kinds of discrete symbols, enabling various mathematical representations of category combinations. Secondly, this can also lead to the idea of abandoning analytic categories altogether. The current trend in NLP is to overcome the need for annotated data, which is insufficient for most languages and tasks. Learning from raw texts and end-to-end models benefit from the enormous amount of raw texts, and challenge supervised NLP models trained on scarse but sophisticated symbolic annotations. But symbols do come back by the window: current research efforts on interpretability of neural models show that NLP without meta-linguistic explicitation is not satisfying. A promising research perspective is to integrate top-down (formalization to data) and bottom-up (from data to neural nets parameters) approaches.

3. Content of the special issue

The special issue contains three papers, two concerning syntactic treebanks and one concerning the evaluation of the quality of annotations. It is striking that the first treebank paper (Abeillé, Clément and Liégeois, *Un corpus arboré pour le français: le French Treebank*) provides an overview and some feedback on the first treebank project for French, namely the French Treebank, which has nourished NLP research for the French language, and whose development spans over the last twenty years. On top of an overview of the major linguistic choices underlying the annotation scheme, this paper is the occasion to provide some feedback on what could have been made differently, in the light of the various uses of the corpus in the last years. The paper presents the first full release of the corpus, namely the complete annotation of the whole corpus, namely meta-data concerning the author and domain of the articles, and, for each sentence, the annotation of multi-word expressions, parts-of-speech, morphological features, syntactic constituents and grammatical functions. Finally, a partial evaluation of the annotation quality is provided for the first time.

Interestingly, the other treebanking paper (Guillaume, de Marneffe and Perrier, *Conversion et amélioration de corpus du français annotés en* Universal Dependencies) focuses on how to have various treebanks converge within the multilingual annotation scheme of the *Universal Dependencies*. The difficulties of retaining linguistic description accuracy while using a scheme meant to be multilingual are described and illustrated with the French case. The paper presents a few annotation choices for cases not fully specified in the UD guidelines. A methodology for improving the quality of the resulting corpora is described, namely the double conversion method (converting into one annotation scheme, and converting back), thanks to the use of graph-rewriting rules. Differences with the original annotation signal errors either in the conversion rules, or in the original annotation.

The paper by Brégeon, Antoine, Villaneau and Halftmeyer, *Redonner du sens à l'accord inter-annotateur : vers une interprétation des mesures d'accord en termes de reproductibilité des annotations* focuses on annotation quality evaluation. Such an evaluation is essential to enable annotated corpora to serve as basis for valuable scientific findings. More precisely, the paper concerns the evaluation of a categorization task, in the multi-annotator setting, in which case the quality evaluation focuses on the inter-annotator agreement. The authors consider that popular chance-corrected measures such as Cohen's kappa and Krippendorff's alpha scores are difficult to interpret, and point out the arbitrary nature of the usual interpretation scale of the kappa. To get a more interpretable measure, the authors propose to evaluate the *stability* of the reference annotation. This is achieved by considering an average variation of the reference that would be obtained when taking a majority vote on subsets of annotators instead of on all annotators (note that this entails that at least three annotators per item are required in order to take subsets).

Experiments conducted both on real and simulated data show a correlation between the multi-annotator kappa score and the proposed reproductibility score, which

the authors consider more interpretable. On a closer look though, the proposed metric varies, for the same kappa value, according to the distribution of divergences, i.e. according to whether the divergences are concentrated on a few items or scattered on many of them. On the former case, the variation rate tends to augment. So, despite its sensitivity to the number of classes and number of annotators to consider in subsets, the metric is proved to provide additional information with respect to the kappa.

Acknowledgements

We are very grateful to the editorial board and to the reviewing committee of the *TAL* journal, with special thanks to Emmanuel Morin.

We would also like to warmly thank the members of this issue's specific reviewing committee: Pascal Amsili (LLF, Université Paris Diderot), Farah Benamara (IRIT, Université Toulouse III - Paul Sabatier), Christophe Benzitoun (ATILF, Université de Lorraine), Delphine Bernhard (LiLPa, Université de Strasbourg), Kim Gerdes (ILPGA, Université Sorbonne Nouvelle), Marie-Catherine de Marneffe (Ohio State University), Paola Merlo (Université de Genève), Thomas François (Université catholique de Louvain), Carlos Ramisch (LIS, Aix-Marseille Université), Benoît Sagot (Almanach, INRIA), Agata Savary (LIFAT, Université de Tours), Djamé Seddah (Almanach, INRIA), Marie Tahon (LIUM, Université du Mans).

4. References

Allén S., "Report on work in computational linguistics at the University of Göteborg", *in* E. Mater, J. Štindlová (eds), *Les Machines dans la linguistique*, Éditions de l'Académie Tchécoslovaque des Sciences, Prague, 1968.

Aston G., Burnard L., *The BNC handbook: exploring the British National Corpus with SARA*, Edinburgh University Press, 1998.

Atkins S., "Tools for computer-aided corpus lexicography: the Hector project", *Acta Linguistica Hungarica*, vol. 41, n° 1-4, p. 5, 1993.

Baker J. K., Stochastic modeling as a means of automatic speech recognition, PhD thesis, Carnegie-Mellon University, 1975.

Baum L. E., Petrie T., Soules G., Weiss N., "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *The Annals of mathematical statistics*, vol. 41, n° 1, p. 164-171, 1970.

Boas F., Hunt G., *Kwakiutl texts*, vol. 5, EJ Brill, 1905.

Bresnan J., "Is syntactic knowledge probabilistic?", *in* S. Featherston, W. Sternefeld (eds), *Roots: Linguistics in Search of Its Evidential Base*, Mouton de Gruyter, 2007.

Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S., "A statistical approach to machine translation", *Computational linguistics*, vol. 16, n° 2, p. 79-85, 1990.

Busa R., *La Terminologia tomistica dell'interiorità (Saggi di metodo per un'interpretazione della metafisica della presenza)*, Archivum Philosophicum Aloisianum, 1949.

Busa R., “Index Thomisticus Sancti Thomae Aquinatis Operum Omnium Indices Et Concordantiae in Quibus Verborum Omnium Et Singulorum Formae Et Lemmata Cum Suis Frequentiis Et Contextibus Variis Modis Referuntur”, 1974.

Chamberlain J., Fort K., Kruschwitz U., Lafourcade M., Poesio M., “Using Games to Create Language Resources”, *in* Gurevych, Iryna, Kim, Jungi (eds), *Theory and Applications of Natural Language Processing*, Springer, 2013.

Church K. W., “A stochastic parts program and noun phrase parser for unrestricted text”, *International Conference on Acoustics, Speech, and Signal Processing*, p. 695-698, 1989.

De Marcken C. G., “Parsing the LOB corpus”, *ACL*, p. 243-251, 1990.

Einarsson J., Talbankens skriftspråkskonkordans, Technical report, Lund University: Department of Scandinavian Languages, 1976a.

Einarsson J., Talbankens talspråkskonkordans, Technical report, Lund University: Department of Scandinavian Languages, 1976b.

Epps P. L., Webster A. K., Woodbury A. C., “A holistic humanities of speaking: Franz Boas and the continuing centrality of texts”, *International Journal of American Linguistics*, vol. 83, n° 1, p. 41-78, 2017.

Fort K., Adda G., Bretonnel Cohen K., “Éthique et traitement automatique des langues et de la parole : entre truismes et tabous”, *Traitement Automatique des Langues*, vol. 57, n° 2, p. 7-19, 2016.

Garside R., Leech G., Sampson G., *The Computational Analysis of English: A Corpus-Based Approach*, Longman, 1987.

Greene B., Rubin G., *Automatic Grammatical Tagging of English*, Department of Linguistics, Brown University, 1971.

Guillaume B., Fort K., Lefèbvre N., “Crowdsourcing Complex Language Resources”, *COLING*, p. 3041-3052, 2016.

Hovy D., Spruit S., Mitchell M., Bender E. M., Strube M., Wallach H., “Proceedings of the First ACL Workshop on Ethics in Natural Language Processing”, Valencia, 2017.

Ide N., Véronis J., “MULTEXT: Multilingual text tools and corpora”, *COLING*, p. 588-592, 1994.

Jones S. E., *Roberto Busa, SJ, and the emergence of humanities computing: the priest and the punched cards*, Routledge, 2016.

Kendall T., “On the history and future of sociolinguistic data”, *Language and Linguistics Compass*, vol. 2, n° 2, p. 332-351, 2008.

Kilgarriff A., “Gold standard datasets for evaluating word sense disambiguation programs”, *Computer Speech & Language*, vol. 12, n° 4, p. 453-472, 1998.

Kučera H., Francis W. N., *Computational analysis of present-day American English*, Brown University Press, 1967.

Lari K., Young S. J., “The estimation of stochastic context-free grammars using the inside-outside algorithm”, *Computer speech & language*, vol. 4, n° 1, p. 35-56, 1990.

Leech G., Garside R., Bryant M., “CLAWS4: the tagging of the British National Corpus”, *The 15th International Conference on Computational Linguistics (COLING 1994)*, 1994.

Liberman M., Wayne C., “Human Language Technology”, *AI Magazine*, 2020.

Liberman M. Y., “The trend towards statistical models in natural language processing”, *Natural Language and Speech*, Springer, p. 1-7, 1991.

MacWhinney B., *The CHILDES project: Tools for analyzing talk, Volume II: The database*, Psychology Press, 2014.

MacWhinney B., Snow C., “The child language data exchange system”, *Journal of child language*, vol. 12, n° 2, p. 271-295, 1985.

Magerman D. M., Marcus M. P., “Parsing a Natural Language Using Mutual Information Statistics.”, *AAAI*, vol. 90, p. 984-989, 1990.

Marcus M. P., Marcinkiewicz M. A., Santorini B., “Building a Large Annotated Corpus of English: The Penn Treebank”, *Computational Linguistics*, 1993.

Mielziner M., *Introduction to the Talmud*, Funk & Wagnalls, 1903.

Moon R., *Words, grammar, text: revisiting the work of John Sinclair*, vol. 18, John Benjamins Publishing, 2009.

Morris W. (ed.), *The American Heritage Dictionary of the English Language*, Houghton-Mifflin, 1969.

Müller F. M. *et al.*, *The First Book of the Hitopadeśa containing the Sanskrit text, with interlinear transliteration, grammatical analysis, and English translation [edited by Max Müller]*, vol. 1, Longman, Green, Longman, Roberts & Green, 1864.

Nerbonne J., “Data-driven dialectology”, *Language and Linguistics Compass*, vol. 3, n° 1, p. 175-198, 2009.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajič J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D., “Universal Dependencies v1: A Multilingual Treebank Collection”, *LREC*, 2016.

Pavlick E., Post M., Irvine A., Kachaev D., Callison-Burch C., “The Language Demographics of Amazon Mechanical Turk”, *Transactions of the Association for Computational Linguistics*, vol. 2, p. 79-92, 2014.

Plank B., Hovy D., Søgaard A., “Linguistically debatable or just plain wrong?”, *ACL*, p. 507-511, 2014.

Shannon C. E., “A mathematical theory of communication”, *Bell system technical journal*, vol. 27, n° 3, p. 379-423, 1948.

Sinclair J. M., *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*, Collins Elt, 1987.

Strong J., *The Exhaustive Concordance of the Bible*, Hodder and Stoughton, 1890.

Taylor A., “Treebanks in Historical Syntax”, *Annual Review of Linguistics*, 2020.

Tesnière L., *Éléments de syntaxe structurale*, Klincksieck, Paris, 1959.

Winchester S., *The Professor and the Madman*, Harper, 1998.

Winter T. N., “Roberto Busa, SJ, and the invention of the machine-generated concordance”, *Faculty Publications, Classics and Religious Studies Department,* p. 70, 1999.