From Close to Distant Reading. Towards the Computational Analysis of "Liber Abbaci"

Letizia Ricci[1], Francesco Grotto[2], Margherita Fantoli[3], Rachele Sprugnoli[4], Marco Passarotti[4], Enrica Salvatori[1], Maria Simi[1]

[1]Università degli Studi di Pisa, Italy, l.ricci29@studenti.unipi.it - {maria.simi,enrica.salvatori}@unipi.it
[2]Scuola Normale Superiore di Pisa, Italy, francesco.grotto1@sns.it
[3]University of Leuven, Belgium, margherita.fantoli@kuleuven.be
[4]Università Cattolica del Sacro Cuore di Milano, Italy, {rachele.sprugnoli,marco.passarotti}@unicatt.it

ABSTRACT

This contribution presents the first steps towards the analysis of Leonardo Fibonacci's Liber Abbaci using computational linguistics methods. The work is currently carried out in the context of a joint research project between the Tuscany Region and the University of Pisa with the help of an interdisciplinary team.

KEYWORDS

Computational Linguistics, annotation, text encoding, Latin.

POSTER

1. INTRODUCTION

Leonardo Fibonacci's *Liber Abbaci* is a weighty medieval treatise on arithmetic and algebra that had a decisive influence in the development of Western mathematics. Traditional reading of the text has never been easy in the past and it is not easy now either. The characteristics of the work have, in fact, delayed its critical print edition until very recently. And, looking at the new editorial format and its price - 17 x 24 cm, cxviii-824 pp. 22 plates f.t. color pp., slipcase, Indian paper, silk binding gilded impressions, € 300 - the volume edited by Enrico Giusti ([2]) certainly is not "manageable" and is clearly addressed to an extremely small niche market. The work itself is full-bodied, complex, and presents some problems in the correct understanding and contextualization of terms related to the world of medieval Mediterranean trade, used by the author in illustrating mathematical problems. For this reason, *Liber Abbaci* has been studied more by mathematicians and historians of Science than by other types of humanists ([5];[7]). In 2018 a joint research project between the Tuscany Region and the University of Pisa (p.i. Pier Daniele Napolitani, University of Pisa) has started with the aim of transforming the critical print edition into a completely searchable digital edition, in order to recover the treasure of linguistic, mathematical and historical information that the work contains and therefore to facilitate the access to its content by different users. Within this overall project, an in-depth study was undertaken on the application of computational linguistics methodologies to the *Liber Abbaci* with the aim of developing systems for the automatic extraction of morphosyntactic and semantic information. Due to the linguistic peculiarities of the text, off-the-shelf tools cannot be used without a considerable loss in the performance (see Sec. 2.1). Thus, manually created high-quality annotated data are needed. In other words, in our work we start from the linguistic annotation of a chapter of the *Liber Abbaci*, relying on digital tools for encoding the results of a critical close reading of the text. As future work, we will develop models based on these data to facilitate a comprehensive approach of such a large-scale masterpiece, that will be interrogated with distant reading methods for the first time.

2. COMPUTATIONAL ANALYSIS

The *Liber Abbaci* is a complex work containing a large variety of topics. While the first 7 chapters discuss mathematical operations, chapters 8-11 have an empirical approach, describing commercial practices and monetary topics. The final chapters (12-15) go back to more abstract mathematical problems. The information contained in the chapters on commercial practices is valuable not only for those interested in the contribution of the work to the history of science, but also as a testimony of the history of economics and trade practices. For this reason, we decided to start our work from chapter VIII using that text for our pilot annotations at both the morphosyntactic and the semantic level. Chapter VIII is made up of 29,858 tokens, corresponding to about 10% of the total length of the book.

2.1 MORPHO-SYNTACTIC ANALYSIS

Beside its scientific interest, *Liber Abbaci* features a very peculiar lexicon, not often represented in the currently available linguistically annotated corpora for Latin. In order to fill this gap, we manually performed tokenization, sentence splitting, Part-of-Speech (PoS) tagging and lemmatization of chapter VIII following the Universal Dependencies framework ([9]). During the annotation, we had to deal with several complex linguistic peculiarities of the text that are typical of Medieval Latin such as monophthongization, the presence of analytical verb forms and a very limited use of enclitics. The greatest difficulties, however, concerned the annotation of units of measurement, names of coins, toponyms and arabisms often not even lemmatized in Medieval Latin dictionaries. The annotation was performed by a master's degree student in Classical languages supported by experts in Latin linguistics and computational linguistics. The Inter-Annotator Agreement was calculated on 30 sentences (1,010 tokens), with the participation of a second scholar, and we registered an almost perfect agreement with a Cohen's kappa of 0.97 for lemmatization and 0.94 for PoS tagging. The resulting dataset is freely available online[1] and has been used to evaluate current available automatic models for the processing of Latin. More specifically, we tested the accuracy of five UDPipe ([12]) models with respect to our gold standard: 1) EvaLatin2020, trained on classical texts in prose released for the EvaLatin evaluation campaign ([11]); 2) ITTB, trained on medieval texts of Thomas Aquinas ([4]); 3) LLCT, trained on Early Medieval charters written in Tuscany ([3]); 4) Proiel, trained on selections classical texts plus the Vulgate New Testament translation ([6]); 5) Perseus, trained on classical texts in prose and poetry ([1]). The scores, reported in Tab. 1, clearly show that current models are not good enough to process the Latin of Fibonacci: indeed, the best participating system at the EvaLatin 2020 achieved an accuracy of 96,2% for lemmatization and 96,7% for PoS tagging on the corresponding test set. The specific domain of the text has a negative impact on both lemmatization and PoS tagging: for example, chapter VIII contains a high frequency of lemmas not present in the training data of the model (>50%). Moreover, not all training data follow the latest version of the Universal Dependencies guidelines (v 2.8) causing some inconsistency of the annotations.

Model	**Lemma**	**PoS**
LLCT	68.8	82.8
EvaLatin2020	63.6	81.9
Perseus	67.5	78.4
ITTB	65.6	77.1
Proiel	60.2	51.6

Table 1. Accuracy of UDPipe models tested on chapter VIII.

The lemmatized text of chapter VIII has been linked to the Knowledge Base of interoperable linguistic resources for Latin developed by the ERC project "Lila: Linking Latin" ([8]). Thanks to the linking, our dataset becomes part of an interoperable ecosystem made of resources of different kinds[2] that can be queried using the SPARQL endpoint of LiLa[3].

2.2 SEMANTIC ANALYSIS

We performed a lexical-semantic analysis of chapter VIII in order to identify and classify single terms and multi-token expressions specific to the domain of trade and commerce, so as to facilitate the search within the text. To do so, we have adopted the UCREL Semantic Analysis tagset[4] ([10]), which provides a set of hierarchical semantic tags. Among those tags, we have decided to select only those relevant to the research objective of the project. From the 21 major discourse fields identified by the original UCREL set, we considered 7 of them, in particular I: money and commerce in industry, M: movement, location, travel and transport, N: numbers and measurement, O: substances, materials, objects and equipment, S: social actions, states and processes, T: time, Z: names and grammar. Each field has specific tags and each tag has an

[1] http://dialogo.di.unipi.it/LiberAbaci/.
[2] https://lila-erc.eu/data/corpora/CorpusFibonacci/id/corpus/Liber%20Abbaci.
[3] https://lila-erc.eu/sparql/.
[4] http://ucrel.lancs.ac.uk/usas/.

identification code with a short definition; for example for coins we have the generic tag I1 Money generally which is divided into I1.1 Money: Affluence, I1.2 Money: Debts, I1.3 Money: Price.
As an annotation tool, we have chosen Catma (see Fig. 1), a flexible and user-friendly online application. Catma allows to work on shared projects and to create a tagset with a hierarchical set of labels. It also provides tools for searching and analyzing the annotated text. In our case we started with the preliminary annotation of chapter VIII of *Liber Abbaci* using the tags briefly mentioned above. The most used tags are: I1 Money generally (frequency 872), with which terms of various types of coins are annotated; N3 Measurement (972), which annotates various units of measurement, the most common being the units of weight; I2 Business (291), which indicates terms referring to commerce; Z2 Geographical names (281) usually occurring with units of measurement or coins, for example *rotuli gerovi* referring to the unit of weight *rotoli* with the value used in Genoa. The text does not contain some tags, such as I3 Work and employment, S2 People, T1 Time, Z1 Personal names.

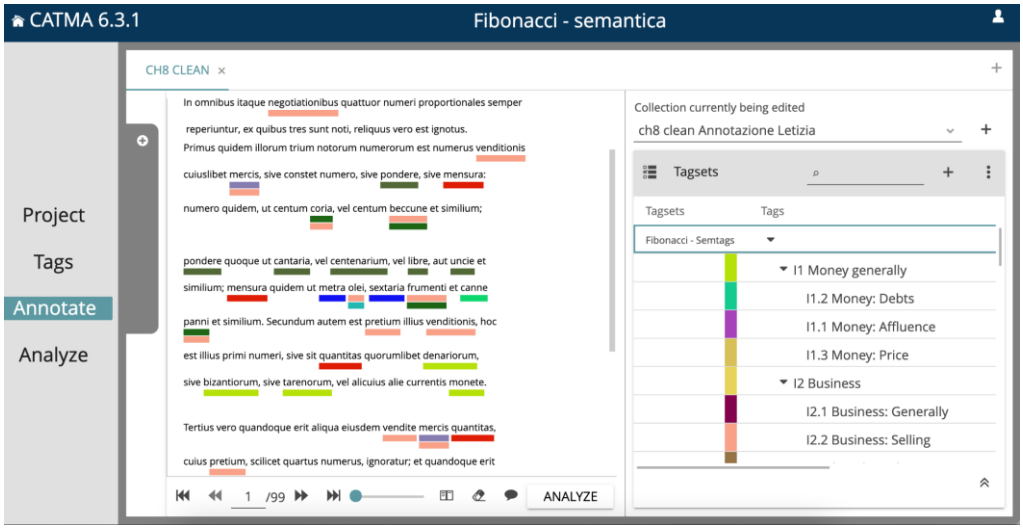

Fig.1 - Catma interface with annotation of chapter VIII.

REFERENCES

[1] Bamman, David, and Gregory Crane. 2011. "The Ancient Greek and Latin Dependency Treebanks." In *Language Technology for Cultural Heritage*, Caroline Sporleder, Antal van den Bosch, Kalliopi Zervanou, 79–98.

[2] Bigolli Pisani, Leonardo vulgo Fibonacci. 2020. *Liber Abbaci*. Edited by Enrico Giusti and Paolo D'Alessandro. Firenze: Olschki.

[3] Cecchini, Flavio Massimiliano, Timo Korkiakangas, and Marco Carlo Passarotti. 2020. "A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages." In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*. Marseille, France: European Language Resources Association (ELRA).

[4] Cecchini, Flavio Massimiliano, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. "Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies." In *Proceedings of the Second Workshop on Universal Dependencies*, 27–36.

[5] Ciocci, Argante, and Enrico Giusti. 2018. "The Twelfth Chapter of Fibonacci's Liber Abaci in Its 1202 Version, Bollettino Di Storia Delle Scienze Matematiche." *Nuncius* 1 (33): 137–39.

[6] Dag, Trygve, Truslew Haug, and Marius L. Jøhndal. 2008. "Creating a Parallel Treebank of the Old Indo-European Bible Translations." In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data*, edited by Caroline Sporleder and Kiril Ribarov, 27–34.

[7] Franci, Raffaella. 2002. "Il Liber Abaci Di Leonardo Fibonacci 1202-2002." *Bollettino Dell'Unione Matematica Italiana* 5 (A.2): 293–328.

[8] Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. "Interlinking through Lemmas. the Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin." *Studi e Saggi Linguistici* 58 (1): 177–212.

[9] Petrov, Slav, Dipanjan Das, and Ryan McDonald. 2012. "Universal Part-of-Speech Tagset." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2089–96. Istanbul, Turkey.

[10] Piao, Scott, Dawn Archer, Olga Mudraya, Paul Rayson, Roger Garside, Tony McEnery, and Andrew Wilson. 2005. "A Large Semantic Lexicon for Corpus Annotation." In *Proceedings from the Corpus Linguistics Conference Series On-Line e-Journal*. Vol. 1. Birmingham, UK.

[11] Sprugnoli, Rachele, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. “Overview of the EvaLatin 2020 Evaluation Campaign.” In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*.

[12] Straka, Milan, and Jana Straková. 2017. “Tokenizing, POS Tagging, Lemmatizing and Parsing Ud 2.0 with Udpipe.” In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. Vancouver, Canada.