LDK 2023

Workflow Reversal and Data Wrangling in Multilingual Diachronic Analysis and Linguistic Linked Open Data Modelling

Florentina Armaselu
University of Luxembourg, Luxembourg, florentina.armaselu@uni.lu

Barbara McGillivray
King's College London
United Kingdom
barbara.mcgillivray@kcl.ac.uk

Chaya Liebeskind
Jerusalem College of Technology
Israel
liebchaya@gmail.com

Giedrė Valūnaitė Oleškevičienė
Mykolas Romeris University, Lithuania
gvalunaite@mruni.eu

Andrius Utka
Magnus University, Lithuania
andrius.utka@vdu.lt

Daniela Gifu
Romanian Academy - Iasi Branch, Romania
daniela.gifu@iit.academiaromana-is.ro

Anas Fahad Khan
Cnr-Istituto di Linguistica Computazionale "Antonio Zampolli", Italy, fahad.khan@ilc.cnr.it

Elena-Simona Apostol
University Politehnica of Bucharest
Romania
elena.apostol@upb.ro

Ciprian-Octavian Truică
University Politehnica of Bucharest
Romania
ciprian.truica@upb.ro

Abstract

The article deals with data wrangling in a multilingual collection intended for diachronic analysis and linguistic linked open data modelling for tracing concept change over time. Two types of static word embeddings are used: word2vec (French and Hebrew data sets), and fastText (Latin and Lithuanian data sets). We model examples from these embeddings via the OntoLex-FrAC formalism. To address the challenge of heterogeneity, we use a minimalist workflow design allowing for both convergence and flexibility in attaining the project goals.

1 Introduction

In data wrangling, the "data required by an application is identified, extracted, cleaned and integrated, to yield a data set that is suitable for exploration and analysis" (Furche et al., 2016, p. 473). The tasks often referred to in this process pertain to data organisation, including data integration and transformation, and data quality, including missing data or anomaly identification (Nazabal et al., 2020). These tasks have also raised questions about the possibilities of automating them (Paton, 2019).

The data wrangling phase described in this proposal is intended to prepare the data for tracing the evolution of concepts in different languages and historical periods through NLP and LLOD approaches. The main challenges of this type of task consist in the heterogeneity of the data sets to be considered for analysis, the need for harmonisation among the different teams involved, and the lack of an established methodology for dealing with the process of data preparation within a multilingual, multi-format, and multi-team context.

Although reported as taking 80% of the data scientist's time (Paton, 2019), data wrangling seems to be less studied so far in digital humanities (DH), and especially in areas that combine natural language processing (NLP), such as diachronic word embeddings, and LLOD representations including spatio-temporal dimensions. Our proposal addresses the question of how to optimise *collaboration* within a DH use case that requires multilingual multi-format corpora (pre-)processing and LLOD modelling by several teams. We approached this question through an adaptation of a method origi-

nated in the domain of engineering, called *workflow reversal* (Chen et al., 2019). It implies an inverse uncertainty propagation and workflow reversal with input-output variable swap to deal with the issue of "handling pre-defined uncertainty associated with design objectives (targets) or constraints (requirements)" (p. 1). We applied the idea in a more general, abstract way, by considering that some requirements and targets can be precisely specified in the workflow, while others can remain underspecified and allow a certain degree of design and implementation flexibility to the different teams.

2 Method

In this section, we present the methodology and the current status of our solution. The main problem was that our data sets varied in many aspects: language, format (TXT, XML; vertical, PoS-tagged, lemmatised), number of files (single, multiple), folder structure (flat, hierarchised), time coverage (ancient, medieval, modern) and genre (Appendix A, Table 1). Although initially we considered unifying all the data formats for the downstream tasks, we realised that this will involve non-trivial preparation and harmonisation work. Finally, for the exploratory design phase, we decided that a certain degree of format variability and independence among the teams can be afforded, provided that a number of common conditions are met at specific points in the processing flow. Therefore, despite the differences in the intermediary steps for our data sets and teams, we were able to define convergence points, through common requirements and outputs in the workflow, that had to be fulfilled for all the involved parts. The main tasks of the workflow were: 1) generate a set of terms and their neighbours resulting from word embedding (word2vec or fastText) and cosine similarity measures; 2) model via OntoLex-FrAC the word embedding results and possibly combine them with dictionary evidence, to represent the evolution of a set of parallel or related concepts in the studied languages.

Figure 1 illustrates the minimal requirements (brace callout) that are demanded by each module or target (rectangular blocks) from the previous modules to accomplish its objectives. Hence, the reversed sense of the arrows, with a left-to-right reading for targets and their needs, and right-to-left, for the actual order of the processing operations. While the types of data wrangling, target tasks, and constraints are specific to our project, we assume that the general method of workflow reversal, understood as a way of identifying the minimal set of specifications and common targets viewed from the reversed perspective of what is needed or intended to be achieved, can be applied to other projects that deal with issues such as the heterogeneity of data and approaches, and multi-team collaboration.

3 Results

Currently, we are in the phase of LLOD modelling, intended to use the OntoLex-FrAC formalism for RDF-based machine-readable dictionaries combined with corpus observables and observations (Chiarcos et al., 2022). The data wrangling and diachronic word embedding tasks included so far experiments with the French, Latin, Hebrew, and Lithuanian data sets. Partial findings from these experiments are expected to be applied to the other corpora from the collection. The data preparation involved different strategies depending on the format characteristics of each data set.

The Lithuanian data set comprised three layers. The representation layer used the original spelling which was transliterated into modern Lithuanian on the next layer, followed by linguistic and morphological annotations. The text was lemmatised and English translations were provided. The decision was to work with the transliteration into the modern Lithuanian layer. Then, the procedures involved extracting text and metadata from XML files and organising the resulting text files by time slice, to prepare them for diachronic word embedding. It was chosen to use FastText, as it is acknowledged to work better for word embeddings in morphologically rich languages, with experimentally proven results in the Lithuanian language (Petkevicius and Vitkute-Adzgauskiene, 2021). The corpus was split into three time periods: 16th, 17th and 18th century. FastText embeddings were generated for each subcorpus for further analysis.

For the Latin corpus, we extracted the publication dates from the metadata available in the corpus file, and normalised the dates so that they were all in a numeric format. This required converting centuries in years or assigning the midpoint between the two extremes in the case of a data range. The input to the embedding training was the lemmatised version of the corpus. We split the corpus into three time intervals: from 450 BCE to 1BCE, from 1CE to 450 CE, and from 451CE to 900 CE. We generated FastText embeddings for each subcor-

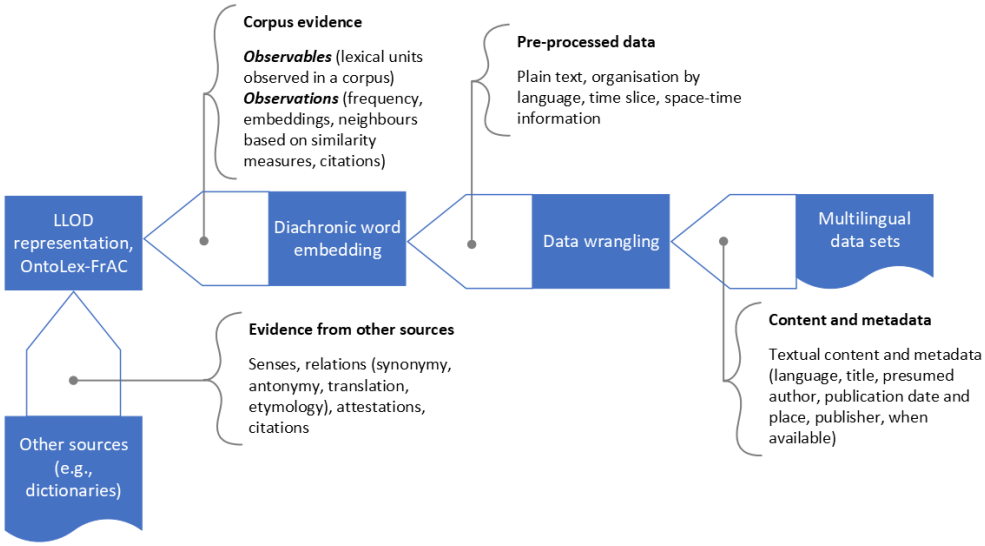

Figure 1: Workflow reversal for multilingual diachronic analysis and LLOD representation

pus, with 100 dimensions, a context window of 5 words to the left and to the right of the target word, and a minimum frequency threshold of 50. In order to make the semantic spaces comparable, we aligned the semantic spaces using the Procrustes Alignment algorithm (Schönemann, 1966).

Minimal pre-processing was performed on the Hebrew Responsa data set before the word embedding (word2vec) phase. Considering the poor performance of a state-of-the-art modern Hebrew POS taggers on the Responsa (Liebeskind et al., 2012), this pre-processing consisted only of white space tokenisation. We split the Responsa into four time intervals: the 11^{th} century until the end of the 15^{th} century, the 16^{th} century, the 17^{th} through the 19^{th} centuries, and the 20^{th} century until today (Liebeskind and Liebeskind, 2020).

The preparation of the Romanian data set included operations such as: acquisition of primary textual data, clearing of copyrights, OCR in some cases, interpretative transliterations in some others, storing, cleaning of data, and metadata completion. From the input DOC and PDF files, raw text was extracted and lists of words were generated. The extracted text was passed to the PoS-tagger that outputs XML files with unknown words marked as NotInDict (Not In Dictionary), i.e., words whose lemmas were not found in the DEXonline lexical database, but also numbers, including years, and proper names. The PoS-tagger included sentence segmentation, tokenisation, and lemmatisation. To create the word embeddings, Radim Rehurek's gensim package, for instance, could be used.

For the BnL Open Data, containing thousands of XML files in a hierarchy of folders and subfolders, an automatic pre-processing was necessary. Figure 3 (Appendix B) illustrates the preparation of the monograph subset (the arrows indicate the input-output direction). The pipeline was produced with KNIME, a software for creating data science workflows. It extracted text and metadata from the BnL hierarchy of folders and XML files, selected only French documents and generated new file names, plain text files, and a new folder structure. The longest horizontal branch (ReadXML to CSV Writer) extracted the textual content from the XML files, and created a flat folder with all the resulting TXT files for French. To the original file names, a prefix was added (language code and publication date from the XML file) to be used in the second KNIME workflow. The three other branches (ending with CSV Writer) produced files for metadata (language, publication date, publisher, persistent ARK identifier), statistics (word and document count by language), and issues (lists of files missing language information). A second KNIME workflow organised the text files by time slice, [1] taking into account elements from the history of Luxembourg, e.g., military and political events, royal decrees and school laws. Other platforms were also tested (OpenRefine and Karma). KNIME was selected since it was open source and dealt well with XML and folder hierarchy processing, and missing data and inconsistency detection.

[1] BnL monographs, time slices: 1690 – 1794; 1795 – 1814; 1815 – 1830; 1831 – 1866; 1867 – 1889; 1890 – 1918.

4 Discussion

For our experiments, we used static word embeddings and gensim word2vec (Rehurek and Sojka, 2010) for French and Hebrew, and fastText (Bojanowski et al., 2017) for Latin and Lithuanian. This required tokenised text, with and without lemmas and PoS, and sentence segmentation. The corpora were structured by time slice (year, decade, century) to determine semantic changes. For each language, we trained our own word embeddings, and we intend to compare the results across language and time period. For example, we were able to qualitatively assess the Latin diachronic embeddings against known instances of lexical semantic change. To mention one such case, the neighbours of the embeddings for the Latin word *pontifex* display evidence of the shift from the domain of the traditional Roman religion (e.g. *sacerdos* 'priest' and *aedes* 'temple' towards terms related to Christianity, such as *missa* 'mass' and *beatus* 'blessed'.

Qualitative assessment was performed for the French data set, after having applied word2vec (5 word window, 100 dimension vectors) by time slice. We compared the list of neighbours resulting from word embedding with dictionary attestations, and found corpus evidence of emerging polysemy within the time period of the data set. For example, we aligned the embedding results of the term *révolution (revolution)* with different senses attested by Ortolang, such as: 'motion of a body around an axis', 'motion of a figure around an axis', 'natural phenomena', and 'political change'. While the attestations always referred to earlier dates than the time intervals of the embeddings, the analysis provided a snapshot of the senses on a timeline and their dictionary-corpus contextualisation.

The word מהפכה (*revolution*) has appeared in numerous contexts throughout the Responsa (as evidenced by its top neighboring terms). The majority of references to revolution in the first era are made in a religious context (כפירה (*atheism*), תשובה (*repentance*)). In the second era, the word is used less frequently. However, it occurs in the context of war and tragedy (אונס (*rape*), הרג (*killing*), מיתה (*death*)) in the third era, which corresponds to the eras in the French corpus, as a consequence of the pogroms that Jews faced during this time. Industrial (מכונות (*machines*), אנרגיה (*energy*)) and medical (החיאה (*resuscitation*), אנאטומיה (*anatomy*)) revolutions, and revolutionary ideological movement (רפורמים (*(Judaism) Reform)*), חילוניות (*secularism*)) pertain to the fourth period.

A qualitative assessment performed on the Lithuanian data set by comparing word embeddings to the dictionary entries revealed that, for example, for the word *ponas (mister, lord)* the polysemy identified in the data set could be attested by the Lithuanian language dictionary[2].

These first results served for exploratory analysis and estimation of the possible outcomes obtained from our data sets, which led us to consider a combination of corpus and lexicographic resources for the subsequent LLOD modelling task. The OntoLex-FrAC model seemed appropriate to it.

No generally agreed upon way of representing diachronic constructs in linked data exists, despite of several proposals within the OntoLex-Lemon framework (see (Armaselu et al., 2022) for a discussion). Currently, we experiment with the Frequency, Attestation, and Corpus information (FrAC) extension of the OntoLex module (McCrae et al., 2017) to represent word embeddings and the relationship between lexical entries and the relevant corpora (Chiarcos et al., 2022), also considering previous work in modelling etymological information in lexical linked data resources (Khan, 2018).

Figure 2 provides a generic example of OntoLex-FrAC combining corpus and dictionary-based attestation for a lexical entry in language l1. This may be connected to other senses, lexical concepts, and entries in other languages through etymological and translation relations. We propose to add a new property and class (`new:dictionary`, `new:Dictionary`) to indicate a dictionary attestation, and a property (`new:neighList`) to store the neighbours in a structured form such as a list. Each neighbour can be represented as an instance of one of the subclasses of `frac:Observable` (lexical entry, lexical sense, form, lexical concept). This type of resource may be used for queries and inferences about semantic change, or enrichment.

The interplay between semantics and pragmatics (e.g., determined by historical, socio-cultural, communication-related factors) should also be considered in representing semantic change and its context. This may involve knowledge- and language-oriented theoretical frameworks, and properties such as `ontolex:usage` for modelling usage and pragmatic nuances of word meaning (Armaselu et al., 2022), or other forms of encoding linguistic

[2]Lietuvių kalbos žodynas (Lithuanian language dictionary, electronic edition). 2017. Vilnius: Lietuvių kalbos insitutas. http://www.lkz.lt (accessed 10 January 2022).

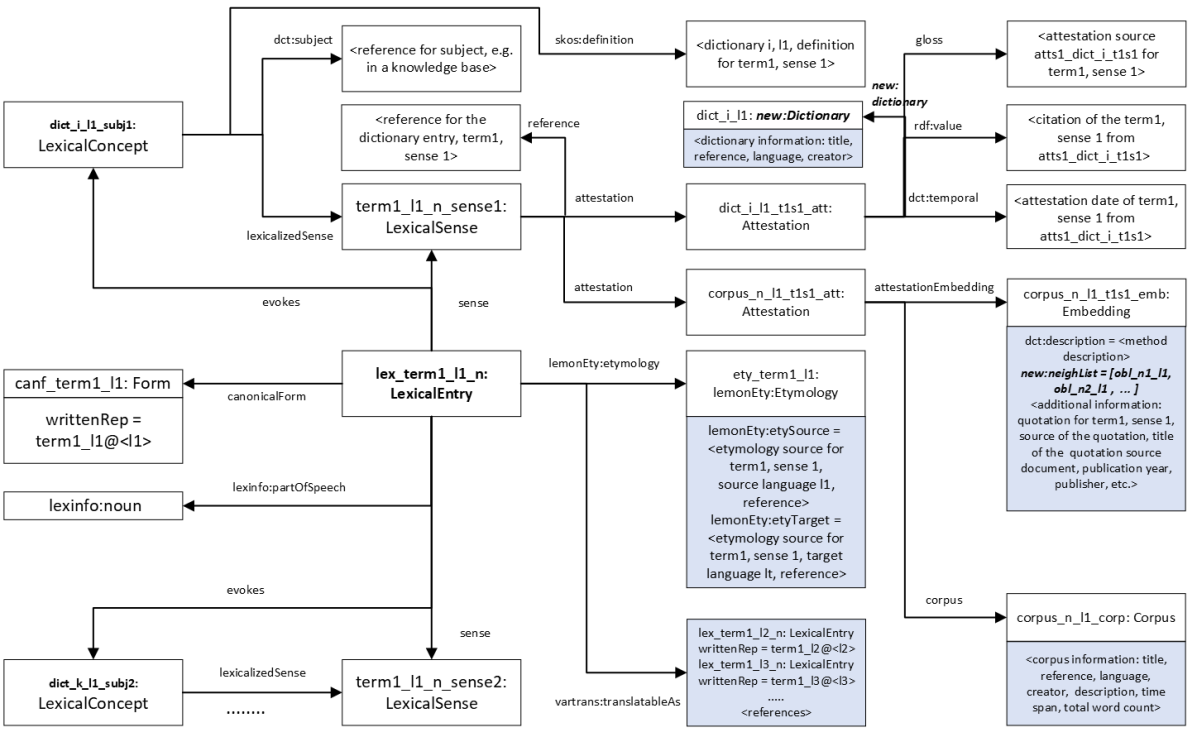

Figure 2: OntoLex-FrAC example combining corpus and dictionary-based attestation (angle brackets: single-item free descriptions; blue-shadowed cells: aggregated descriptions)

content as LLOD still under investigation (Bosque-Gil et al., 2018; Gromann et al., 2022).

5 Conclusion

The proposal focuses on data wrangling in multi-language data sets with various sizes, formats, time spans, and downstream tasks. We argue that a combination of NLP methods and LLOD formalisms, such as diachronic word embedding and OntoLex-FrAC, as well as corpus- and lexicographic-based evidence, can serve in creating inter-operable and more context-rich LLOD resources for detecting and representing semantic change.

We applied the concept of workflow reversal as a general framework for devising a common yet flexible roadmap for our data preparation phase. We defined a minimal set of functional blocks and requirements necessary to accomplish the intended tasks and allowed a certain degree of freedom in their implementation, according to the specificity of each data set, language, and team. The main challenge in applying this type of method may consists in finding a balance between the under-specified and the well-defined parts of the workflow, and avoiding downstream divergence that can impede the project goals. We will use this exploratory design phase to refine and apply the implementation requirements to each language, with the aim of building a multilingual sample of interconnected LLOD diachronic ontologies. Since some of the data sets were rather limited in time coverage, it may be envisaged to complement them, for instance by using multilingual corpora available online via repositories such as Wikimedia Downloads.

Acknowledgment

This article is based upon work from COST Action *Nexus Linguarum, European network for Web-centred linguistic data science*, supported by COST (European Cooperation in Science and Technology). www.cost.eu.

Authors' contribution

F.A. wrote the manuscript and contributed to the methodological design, data processing and analysis for French and LLOD modelling; BMcG conducted the processing and analysis of the Latin data, contributed to the methodological design and wrote the parts of sections 3, 4 and Table 1 relative to Latin; C.L. conducted the processing and analysis of the Hebrew data and contributed to sections 3, 4 and Table 1 relative to Hebrew; G.V.O.

and A.U. conducted the processing and analysis of the Lithuanian data and contributed to sections 3, 4 and Table 1 relative to Lithuanian; D.G. conducted the processing of the Romanian data and contributed to sections 1, 3 and Table 1 relative to Romanian; A.F.K. contributed to discussions on the modelling of the results as LLOD in section 4; E.S.A. contributed to sections 1, 3 and 5; C.O.T. contributed to sections 2 and 3. All authors reviewed the final manuscript.

References

Florentina Armaselu, Elena-Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Giedrė Valūnaitė Oleškevičienė, and Marieke van Erp. 2022. LL(O)D and NLP perspectives on semantic change for humanities research. *Semantic Web*, 13(6):1051–1080.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, and Adrián Gómez-Pérez. 2018. Models to represent linguistic linked data. *Natural Language Engineering*, 24(6):811–859.

Xin Chen, Arturo Molina-Cristóbal, Marin D. Guenov, Varun C. Datta, and Atif Riaz. 2019. *A Novel Method for Inverse Uncertainty Propagation*, volume 48 of *Computational Methods in Applied Sciences*, pages 353–370.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC. In *International Conference on Computational Linguistics*, pages 4018–4027.

Tim Furche, Georg Gottlob, and Leonid Libkin. 2016. Data wrangling for big data: Challenges and opportunities. In *International Conference on Extending Database Technology*, pages 473–478.

Jolanta Gelumbeckaite, Mindaugas Šinkunas, and Vytautas Zinkevicius. 2012. Old Lithuanian reference corpus (SLIEKKAS and automated grammatical annotation). *Journal for Language Technology and Computational Linguistics*, 27(2):83–96.

Daniela Gifu. 2016. *Lexical Semantics in Text Processing. Contrastive Diachronic Studies on Romanian Language*. PhD thesis, "Alexandru Ioan Cuza" Uni-versity of Iași, Romania.

Dagmar Gromann, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Verginica Mititelu, Liudmila Mockiene, Michael Rosner, Ineke Schuurman, Gilles Sérasset, Purificação Silvano, Blerina Spahiu, Ciprian-Octavian Truică, Andrius Utka, and Giedre Valunaite Oleskeviciene. 2022. Multilinguality and LLOD: A survey across linguistic description levels. *Semantic Web 1 (0) 1–39, IOS Press (currently under review)*.

Anas Khan. 2018. Towards the representation of etymological data on the Semantic Web. *Information*, 9(12):304.

Chaya Liebeskind, Ido Dagan, and Jonathan Schler. 2012. Statistical thesaurus construction for a morphologically rich language. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 59–64.

Chaya Liebeskind and Shmuel Liebeskind. 2020. Deep learning for period classification of historical Hebrew texts. *Journal of Data Mining & Digital Humanities*, 2020:5864.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, and Paul Buitelaar. 2017. The OntoLex-Lemon model: development and applications. In *eLex 2017 Conference*, pages 587–597.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. *New Methods in Historical Corpus Linguistics, Tübingen: Narr*, pages 10.

Alfredo Nazabal, Christopher K. I. Williams, Giovanni Colavizza, Camila Rangel Smith, and Angus Williams. 2020. Data engineering for data analytics: A classification of the issues, and case studies. *arXiv:2004.12929 [cs]*. ArXiv: 2004.12929.

Norman W. Paton. 2019. Automating data preparation: Can we? Should we? Must we? In *International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data*, pages 1–5.

Mindaugas Petkevicius and Daiva Vitkute-Adzgauskiene. 2021. Intrinsic word embedding model evaluation for Lithuanian language using adapted similarity and relatedness benchmark datasets. In *IVUS*, pages 122–131.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *LREC 2010 Workshop New Challenges for NLP Frameworks*, pages 45–50.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10.

Alessandro Vatri and Barbara McGillivray. 2018. The Diorisis Ancient Greek corpus: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55–65.

Appendix A. Datasets[3]

Data set	Language	Time span	Size	Format	Genre
LatinISE	Latin	2nd c. BCE - 20th c. CE	10 mil. word tokens	TXT, vertical format, lemmatised, PoS-tagged	Literature, philosophy, law, religion, technical writings, letters
Diorisis	Ancient Greek	8th c. BCE - 5th c. CE	10,206,421 word tokens	TXT, enriched with morphological information, lemmatised, PoS-tagged	Literature, philosophy, historiography, scriptures, technical writings, letters
RODICA	Romanian	19th c. (second decade)	over 5 mil. lexical tokens	TXT, XML, PoS-tagged, lemmatised	Newspapers from Moldavia, Wallachia, Transylvania and Bessarabia
SLIEKKAS	Old Lithuanian	16th - 18th c.	10 texts, 350,000 words	TXT, representation layer (old alphabet); transliterated layer (modern Lithuanian alphabet); linguistic and morphological annotations; lemmatised; English translations	Prose and poetry, religious texts (prayers, catechisms, hymnals and sermons)
BnL Open Data	French, German, Luxembourgish	1690-1918 (monographs); 1841-1878 (newspapers)	23,663 newspaper issues, 510,505 articles; 504 monographs, 33,477 chapters	XML, Dublin Core	Monographs: literature, history, philosophy, geography, religion; newspapers
Responsa	Hebrew	11th -21st c.	76,710 articles, about 100 mil. word tokens	TXT	Questions and rabbinic answers on daily issues (law, health, commerce, marriage, education, Jewish customs)

Table 1: Description of the data sets

Appendix B. KNIME workflow

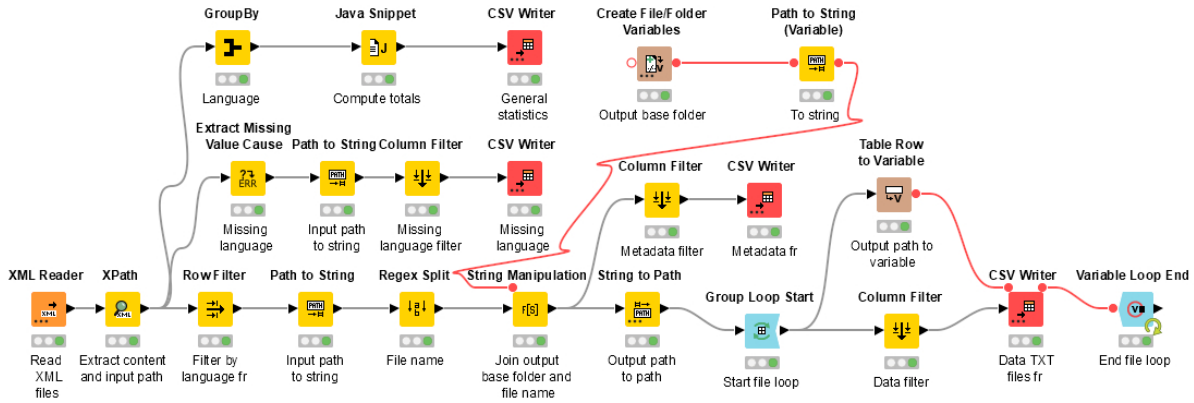

Figure 3: KNIME workflow for the preparation of the BnL Open Data set (French monographs)

[3] LatinISE (McGillivray and Kilgarriff, 2013); Diorisis (Vatri and McGillivray, 2018); RODICA (ROmanian DIachonic Corpus with Annotations) (Gifu, 2016); SLIEKKAS (Gelumbeckaite et al., 2012); Bibliothèque nationale du Luxembourg, BnL Open Data; Responsa (Liebeskind and Liebeskind, 2020).