Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants

Jean-Baptiste Camps, Elena Spadini, Lucence Ing

Collating Medieval Vernacular Texts: Aligning Witnesses, Classifying Variants

Jean-Baptiste Camps (Jean-Baptiste.Camps@chartes.psl.eu), École nationale des chartes (PSL)

Lucence Ing (lucence.ing@chartes.psl.eu), École nationale des chartes (PSL)

Elena Spadini (elena.spadini@unil.ch), Université de Lausanne, Switzerland

Introduction

Aligning different versions of the same work is both a computational and a philological challenge. In particular, the collation of witnesses of an ancient or medieval text poses specific difficulties due to the coexistence of macro-structural and localised variants, including a large number of formal variants.

We present an experimental computer-assisted workflow for aligning several witnesses and classifying variants. Formal and substantive variants are examples of categories especially relevant for languages which are unstable in their graphic system, as are medieval languages. The case studies are in Old French, and, marginally, Old Spanish.

The distinction between formal and substantive variants enables to treat them separately. Stemmatology, for instance, will be mostly interested in the former (even if this has been challenged in Andrews, 2016), while, for linguistic analysis the latter are needed. In automatic collation, based on full transcription of the texts to be compared, the formal variation is generally preserved, but temporarily nullified by means of normalisation or fuzzy match: this enables an accurate alignment of the texts and at the same time the preservation of the original forms.

How to handle variation

Medieval texts, especially in vernacular, often exhibit important variation. At the phrases or words levels, syntactic or graphic variations account for diachronic and diatopic differences, varying scribal practices and the plurality of graphematic standards. This makes it difficult to align sequences between texts, when they have very few letters in common, e.g., *Cait del fuere* | *Chiet dou fuerre* | *Kiet du feurre* (‘[The sword] falls of the scabbard’).

Difficulties due to spelling or flexional variation only add up to already existing variations in word order or substance. Consider the following example taken from Chrétien de Troye's *Chevalier au lion* (Meyer, 2006, v. 3701):

> *H* Li frans li dolz ou ert il donques
>
> *P* Li frans li dous ou estoit donques
>
> *V* Li franz li doz ou ert il donques
>
> *F* Li frans li dols ou ert il donques

G Li biaux li preuz ou estoit donques

A Li preus li frans u est il donques

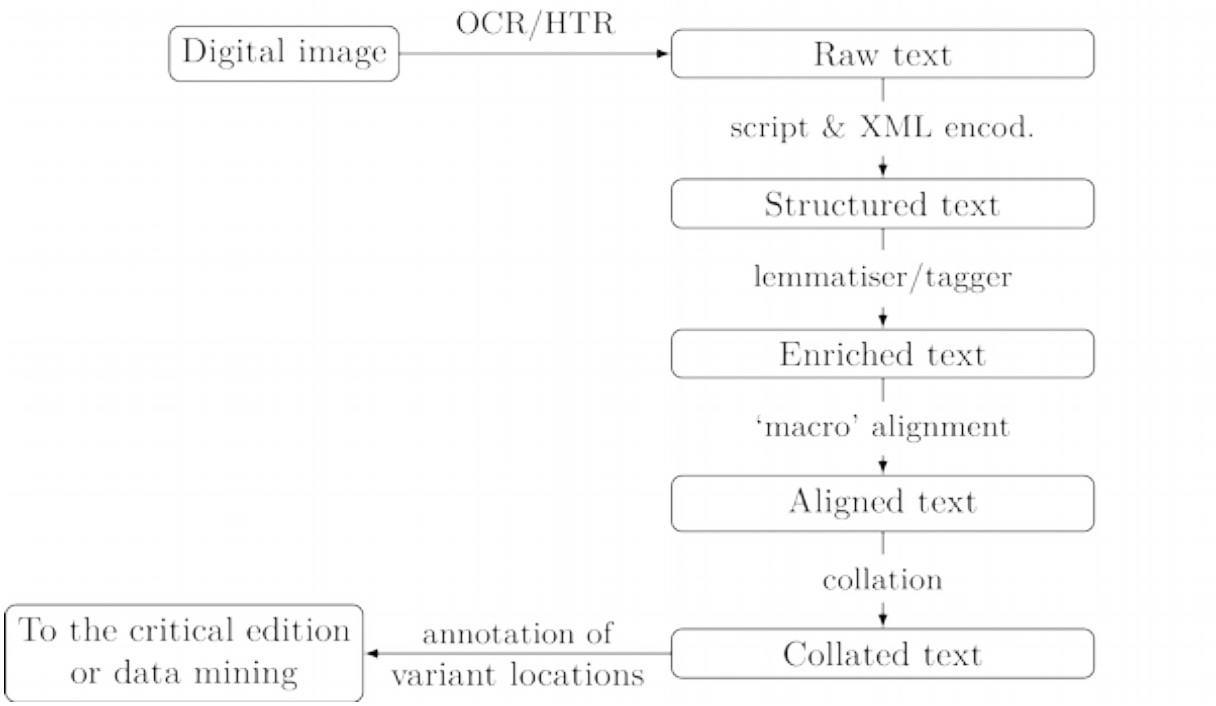
S Li preus li frans u ert il donques

R Li frans li dols u ert il donques

M Li frans li preus ou est il donques

Spelling (e.g., *dolz*, *dous*, *doz*, *dols*) and flexional variants (*est*, *ert*, *estoit*) go along with substitutions (*dous* | *preus* or *biaux* | *frans*), additions/deletions (*il*), or permutations (*preuz*). In such a case, clearing out spelling and flexional variation might help in resolving the other difficulties.

This paper offers a new approach to the normalisation task made possible by the developments in the field of NLP and the resources now available for medieval languages, following the steps described in fig. 1.



Processing workflow

The initial step is the acquisition of the text, from the digital image, done by a combination of manual transcription (for producing ground truth), automated handwritten text recognition, and post-correction. The raw text thus obtained is then structured and stored in an XML/TEI based format. All these tasks are performed before the normalisation step, here represented by lemmatization and linguistic annotation, done with the help of neural network-based taggers/lemmatizers.

Traditionally, normalisation consists of the preparation of the texts for alignment and might imply lowercasing, removing punctuation or editorial markup, as well as the temporary removal of formal features (Silva and Love, 1969 ; Robinson, 1989). Our proposal is to move to an automatic normalisation performed using NLP tools. Each token (i.e. word) is annotated with linguistic information such as part of speech, lemma and detailed morphological information. This kind of normalisation is only possible when suitable resources are available. For Old Spanish, Freeling (Padró Stanilovsky, 2012) provides a specific module (Boleda, 2011; Porta et al., 2013). For Old French, we used the data provided by the *Geste* corpus (Camps et al., 2016), annotated with lemmas, as well as POS and morph tags according to the Cattex scheme (Prevost et al., 2013). With this data, we trained a neural tagger/lemmatizer suitable for variation-rich languages (Kestemont et al., 2017 ; Manjavacas et al., 2019). On the test set, accuracy reached 94.5 and 95% for lemmatization and POS-tagging, and was in the range 94-98.5% for different morphological features.

After normalisation, the texts enriched with linguistic information can be used to perform the alignment. Variation in structure, order or content in medieval texts is favoured by the existence of ‘active textual transmission’ (Vàrvaro, 1970) and by processes of rewriting, prosification/versification, etc. Changes in the order of the structural entities (verses, paragraphs, etc.) are also common. In order to collate these displaced entities, a phase of macro-structural alignment might be needed. This process can be done by a combination of direct expertise and tools conceived for detecting paraphrase, text reuse or computing similarities (Büchler et al., 2014; Jänicke and Wrisley, 2018).

The very collation is then made by using the collation program CollateX (Dekker et al., 2011 and 2015) in its Python version. CollateX uses multiple alignment algorithms, suitable for the comparison of more than two witnesses (Spadini, 2017); its modular structure, based on the Gothenburg model, enables the user to intervene on each module separately and to add new ones.

Automatic categorization of variants

All these software bricks can be integrated in a more complex pipeline up to the the final output. The modular structure of CollateX enables us to adjust the alignment and the visualization phases, in order to take into account the linguistic annotations added to each token. The alignment is performed directly on the annotation, used as a normalised form. In the creation of the output, some rules are added to compare the original forms with the annotation and to assign a category to the variant. For example, the category 'formal variant' is assigned to aligned tokens which have the same annotations but different original forms, such as:

mielz (pos: adverb; lemma: *mieus*),

miels (pos: adverb; lemma: *mieus*),

miaus (pos: adverb; lemma: *mieus*).

Additional rules can be used for classifying variants into finer-grained categories, using linguistic annotation (fig. 2).

lemma	POS	morph	form	category	Ex. subcat.	Ex. cases
=	=	=	≠	graphematic	diatopic diachronic	*chivalier* | *chevalier* *estet* | *esté*
=	=	≠	≠	flexional	verbal nominal	*est* | *estoient* *bar* | *baron*
=	≠	≠	≠	morphosyntactic	nominalization	*aimer* | *(li) aimers*
≠	≠	≠	≠	lexical	derivational paradigmatic semantic	*creanter* | *acreanter* *chat* | *chien* *chevalier* | *charete*

Possible classification of variants using linguistic annotation, with examples of possible subcategories and cases. The broad paradigmatic subcategory encompasses synonyms, cohyponyms, hypero-/hyponymes or holo-/meronymes; the semantic subcategory is reserved for lexemes who do not hold this type of relation between them.

Conclusions and Further research

This paper presents some early results of an ongoing research on automatic collation and categorization of variants. Performing normalization using NLP tools not only speeds up the task, but also makes the identification of fine-grained categories possible. The case studies show the strong and weak points of this proposal and of the technical solutions for its implementation. Eventually, this research forces us to reflect upon the importance of having software components which are open and modular, in order to improve them and to include them in computational pipelines.

References

Andrews, T. L. (2016). Analysis of Variation Significance in Artificial Traditions Using Stemmaweb. *Digital Scholarship in the Humanities*, **31**(3): 523–39 doi:10.1093/llc/fqu072.

Boleda, G. (2011). Extending the tool, or how to annotate historical language varieties. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. pp. 1–9.

Büchler, M., Burns, P. R., Müller, M., Franzini, E. and Franzini, G. (2014). Towards a Historical Text Re-use Detection. In Biemann, C. and Mehler, A. (eds), *Text Mining: From Ontology Learning to Automated Text Processing Applications*. (Theory and Applications of Natural Language Processing).

Cham: Springer International Publishing, pp. 221–38 doi:10.1007/978-3-319-12655-5_11. https://doi.org/10.1007/978-3-319-12655-5_11 (accessed 25 April 2019).

Camps, J.-B., Albarran, E., Cochet, A. and Ing, L. (2016). *Jean-Baptiste-Camps/Geste: Geste: Un Corpus de Chansons de Geste, 2016-…*. Zenodo doi:10.5281/zenodo.2630574. https://zenodo.org/record/2630574 (accessed 25 April 2019).

Haentjens Dekker, R., Hulle, D. van, Middell, G., Neyt, V. and Zundert, J. van (2015). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities*, **30**(3): 452–70 doi:10.1093/llc/fqu007.

Haentjens Dekker, R. and Middell, G. (2010). *CollateX*. http://collatex.net/ (accessed 25 April 2019).

Jänicke, S. and Wrisley, D. J. (2017). Visualizing Mouvance: Toward a visual analysis of variant medieval text traditions. *Digital Scholarship in the Humanities*, **32**(suppl_2): ii106–23 doi:10.1093/llc/fqx033.

Kestemont, M., Pauw, G. de, Nie, R. van and Daelemans, W. (2017). Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities*, **32**(4): 797–815 doi:10.1093/llc/fqw034.

Manjavacas, E., Kádár, Á. and Kestemont, Mike, M. (2019). Improving Lemmatization of Non-Standard Languages with Joint Learning. arXiv preprint arXiv:1903.06939 (accessed 25 April 2019).

Meyer, K. (2006). *Transcription Synoptique Des Manuscrits et Fragments Du Chevalier Au Lion Par Chrétien de Troyes*. Université d'Ottawa: Faculté des Arts, Laboratoire de français ancien http://francaisancien.net/activites/textes/kmeyer/kpres.html (accessed 25 April 2019).

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May, 2012.*

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. San Rafael: Morgan and Claypool.

Porta, J., Sancho, J.-L. and Gómez, J. (2013). Edit transducers for spelling variation in Old Spanish. *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*. Linköping University Electronic Press, pp. 70–79.

Prévost, S., Guillot, C., Lavrentiev, A. and Heiden, S. (2013). *Jeu d'étiquettes Morphosyntaxiques CATTEX2009*. version 2.0 http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf (accessed 25 April 2019).

Robinson, P. M. W. (1989). The Collation and Textual Criticism of Icelandic Manuscripts (1): Collation. *Literary and Linguistic Computing*, **4**(2): 99–105 doi:10.1093/llc/4.2.99.

Silva, G. and Love, H. (1969). The identification of text variants by computer. *Information Storage and Retrieval*, **5**(3): 89–108 doi:10.1016/0020-0271(69)90014-X.

Spadini, E. (2017). The role of the base manuscript in the collation of medieval texts. In Boot, P. and alii (eds), *Advances in Digital Scholarly Editing. Papers Presented at the DiXiT Conferences in The Hague, Cologne, and Antwerp*. Leiden: Sidestone Press, pp. 345–50.

Varvaro, A. (1970). Critica dei testi classica e romanza. Problemi comuni ed esperienze diverse. *Rendiconti Dell'Accademia Di Archeologia, Lettere e Belle Arti Di Napoli*, **XLV**: 73–117.