Historical Corpus and Historical Dictionary: Merging Two Ongoing Projects of Old French by Integrating their Editing Systems

Sabine Tittel
Heidelberg Academy of Sciences and Humanities
E-mail: sabine.tittel@urz.uni-heidelberg.de

Abstract

To combine corpus data with dictionary data has two advantages: (i) It embeds the vocabulary of the corpus texts within the overall system of the language, and it semantically disambiguates the texts. (ii) The corpus data enrich the dictionary and shed new light on the comprehension of the vocabulary. The retrospective integration of corpus data into a dictionary is a task that has to focus on two aspects, (i) on the integration of the word forms, and (ii) on the semantic integration of the words. This second aspect continues to be an important issue, particularly for historical languages. Automated solutions do not exist. In this paper, we present the retrospective integration – both with a graphical and a semantic focus – of the corpus of Old French legal texts, Documents linguistiques galloromans (with approx. 800,000 attestations of Old French lexemes), into the Dictionnaire étymologique de l’ancien français (with 83,000 dictionary entries). We have implemented a semi-automated process resulting in a time-saving editorial workflow to accomplish the data integration. Further, we have created a twofold publication concept for the dictionary entries that makes for a straightforward way of enriching the dictionary with the valuable material of the domain of Old French law.

Keywords: historical lexicography, corpus linguistics, Old French, dictionary writing system, scholarly digital text edition, history of law

1 Introduction

The retrospective integration of two large and long-standing projects of Medieval French, i.e., the corpus of Old French legal documents, Documents linguistiques galloromans – DocLing, and the comprehensive dictionary of the Old French language, Dictionnaire étymologique de l’ancien français – DEAF, is a challenging task. Such a retrospective approach cannot profit from the advantages that a newly created corpus lexicographic venture has. The latter typically defines its linguistic corpus in the initial phase of the project as a well-integrated part of the system architecture. An example in the field of Old French is the Dictionnaire Électronique de Chrétien de Troyes – DÉCT (Kunstmann 2007–2014).

Instead, combining the two long-established projects DocLing and DEAF has to deal with distinct data formats that are specific for the corpus and the dictionary, respectively, and with the adaptation of a tailor-made electronic dictionary writing system.

In this paper, we present our data integration that focuses on two main aspects. Firstly, we have implemented the integration with respect to the word forms attested in the corpus texts. This means that we merge the approximately 800,000 word occurrences of DocLing within the parts of the approximately 83,000 DEAF entries that present the graphical realizations of the lexemes. Secondly, we have implemented the integration with respect to the meaning of the words. This means that we perform a semantic mapping of the word occurrences within the DEAF entries. This second aspect is not an obvious task and continues to be a challenging issue, particularly for historical languages.
From the point of view of historical linguistics, the data integration has two main benefits: (i) It enriches the dictionary with data from the discourse tradition of medieval law that had previously been widely unnoticed by lexicography. (ii) It embeds the vocabulary of the corpus texts within the overall system of the medieval French language. At the same time, it creates a means for the semantic disambiguation of the vocabulary and, thus, for the understanding of its meaning. We believe our approach is promising for the integration of other corpora and dictionaries.

This paper is structured as follows: In Section 2, we introduce the two lexical resources DEAF and DocLing. Section 3 explains our integration approach, and Section 4 presents the semi-automated workflow of the data integration: First, we discuss the mapping of the data models and the rules for the data import and export. Then, we show the steps of the semi-automated workflow and the graphical user interfaces we developed for the integration. Section 5 presents the online publication of the integrated resources with two successive release steps. Section 6 discusses the added values for both the dictionary and for the corpus from the point of view of the historical content. In Section 7, we address the remaining issue of how to link back from the corpus to the dictionary, and Section 6 concludes our work.

2 The Lexical Resources

The DEAF (Baldinger, Möhren & Städtler 1971–) is a longstanding dictionary compiled under the aegis of the Heidelberg Academy of Sciences and Humanities in Heidelberg, Germany. It researches the Old French language from its first resource 842 AD until ca. 1350. The dictionary is traditionally published as a series of printed books. Since 2010, it is also published as a versatile electronic version with online dictionary entries and elaborate research functions, called DEAF électronique (DEAFél).¹ The DEAF organizes the Old French lexicon in word families to show the etymological relations between single lexemes. The main-lemma of a dictionary entry is the lexeme that is developed or borrowed from a Latin, Greek, etc., origin. The derivations from this lexeme are the sub-lemmata.

The online publication DEAFél consists of two parts: DEAFpré and DEAFplus. DEAFplus is the online version of the well-known dictionary DEAF; it consists of extensive articles of the scientifically acknowledged lexicographical quality that has characterized the DEAF for more than 30 years (DEAFplus features a number of added values compared to the printed book and explaining the ‘plus’, cf. Tittel 2010). DEAFpré is not a dictionary in its proper sense, but offers the complete raw material of the DEAF. This material is accessible online in the form of compendious articles that are orthographically and semantically structured in a semi-automated manner. Therefore, it is valuable for all research within our discipline as long as DEAFplus does not cover the entire alphabet. Together, DEAFplus and DEAFpré form DEAFél with approximately 83,000 entries.²

The dictionary draws its source material from an open textual corpus. This corpus consists of three components: (i) the entirety of accessible scholarly text editions of Old French texts (currently around 3,000 primary texts within around 10,000 manuscripts), (ii) the information published in the secondary literature (monographies, journals), and (iii) the information published in related dictionaries. The 1.5 million handwritten and now digitized fiches (slips) lead to 12 million attestations within the corpus. For the most part, the textual corpus of the dictionary cannot be digitally accessed. Even though

¹ See https://deaf-server.adw.uni-heidelberg.de [accessed 03-28-2018].
² The lemmata treated in the form of DEAFplus (letters D–K) will add up to approximately 9,000 in 2020; approximately 73,000 lemmata will remain as DEAFpré (the rest of the alphabet) for the time being. The division of the dictionary into two considerably different parts has been implemented in 2010; it is due to changed monetary conditions affecting the duration of the project.
the editorial process does to some extent integrate corpus queries on a few digital texts, we can say that the DEAF is clearly not a corpus lexicographic endeavor.

In 2014, the DEAF started a cooperation with DocLing (University of Zurich, Glessgen 1998–). DocLing is one of the most significant projects of Old French corpus linguistics. It comprises digital scholarly text editions of 2,185 Medieval French charters (deeds of donation, contracts of purchase, inheritance matter, etc.) dating between 1205 AD and ca. 1450 and with approximately 800,000 word-occurrences. These text editions were created within the framework of the corpus project. They make accessible the important textual genre of legal documents, and are textual witnesses that cover all aspects of human social interaction. For the time being, the DocLing material will be integrated into the part of the DEAF that we call DEAFpré.

3 Integration Approach

The aim of the DocLing-DEAF cooperation is the full integration of the DocLing data into the DEAF dictionary. The attestations of Old French lexemes from the text editions of DocLing shall find their correct place within the dictionary entries of the DEAF.

According to Asmussen (2013), there exist two prototypical approaches to the retrospective integration of a corpus and a dictionary. The first is to add “deliberately selected and processed text material from a corpus […] to a dictionary to give more citations for each definition in the dictionary” (Asmussen 2013: 1084). He qualifies this approach as being a tedious and error-prone task that should not be carried out (ibid.: 1087). Instead, he promotes the second approach, i.e., to establish a virtual interlinking based on orthographical and morphological matches of the lemmatized corpus data with dictionary entries. We consider this second approach insufficient for the following reason: The interlinking focuses only on the graphical realizations of the lexemes. The question of how to establish pointers from the lexical units to the right sense within a dictionary entry is thus not resolved. It is clear that this approach still needs the semantic disambiguation of the corpus data and a correct semantic mapping. As such, the error-prone task remains. Asmussen identifies the question of how to semantically map the corpus data as an important issue for future research (ibid.). We will present our solution to this question in this paper.

With the integration, we follow the second approach while we also address the issue of how to perform the semantic mapping. Hence, we have two objectives. The first objective is to integrate the corpus data with respect to the graphical realizations of the lexemes, i.e., within the apparatus of graphical variants of the respective dictionary articles. The second objective – and this is our main concern – is to integrate the corpus data with respect to the meaning of the words, i.e., within the

3 See http://www.rose.uzh.ch/docling/ [accessed 03-28-2018].
4 DocLing also comprises documents of the Romance speaking Suisse regions and the Francoprovençal linguistic area. For the moment, these are not relevant for our purposes.
5 Most other corpora with texts of Old French incorporate already existing text editions, e.g. the Textes de Français Ancien – TFA (Kunstmann, Ottawa, http://arlill-project.uchicago.edu/content/tfa [accessed 03-27-2018]), the Base de Français Médieval – BFM (Guillot / Heiden / Lavrentiev / Marchello-Nizia / Prévost, Lyon, http://bfm.ens-lyon.fr [accessed 03-27-2018]) and the Corpus représentatif des premiers textes français – CoRPTEF (Guillot, Lyon, http://corptef.ens-lyon.fr [accessed 03-27-2018]), the editions of the two oldest texts have been redone for this purpose, i.e., the Serments de Strasbourg, 842 AD, and the Séquence de sainte Eulalie, ca. 900 AD).
6 Similar to the medieval stage of other Romance languages, Old French does not have a consistent orthographic norm. Each scribe of a manuscript realized the sound of a word in his own fashion, influenced both by random circumstances and by his dialect that could differ significantly from what we consider the standardized Old French language. Thus, we find a great variety of spellings for the same word that we assign, as graphical variants, to the canonical form that is the lemma of the dictionary entry (cf. Möhren 2015).
semantic tree of the respective dictionary entries. To achieve this, we have established a workflow for the graphical integration as well as for the semantic mapping to the corresponding sense.

We merge the corpus and the dictionary data – graphically as well as semantically – based on the following assumption: The DEAF dictionary entries constitute a lightweight ontology. Note that this ontology is characterized by a very low degree of axiomatization as opposed to other, more formalized types of ontologies within the ontology spectrum, such as taxonomies and logical languages (cf. Grimm et al. 2011: 522–525). In this ontology, we understand the dictionary entries, their hierarchical organization into main- and sub-lemmata, and their respective semantic trees of main-senses and sub-senses as entities. This ontology constitutes the framework for the integration of the corpus data: We assign the lexical units of DocLing to the concepts of the ontology. These entities are represented in the data format.

4 Integration with Semi-Automated Workflow

The integration of the DocLing data into the dictionary is carried out on both the level of the back-end and on the front-end, i.e., on the level of the applications and also on the level of their graphical user interfaces (GUIs). In the following, we refer to the DocLing application as Phoenix2. The DEAF application is the tailor-made dictionary writing system, in the following referred to as DEAF-DWS.

4.1 Mapping of the Data Models

On the level of the back-end we have implemented a bidirectional data exchange: Each word occurrence plus a specific set of metadata is imported from Phoenix2 into DEAF-DWS, edited there and then written back to Phoenix2.

Naturally, the data models differ between the two systems. The data models of Phoenix2 and DEAF-DWS represent the structure specific to the textual domain of DocLing and the lexicographical domain of the DEAF, respectively. Fortunately, the parts of the data models relevant to the integration match conceptually to a large degree: The basic data entity to model an attestation of a given lexeme in Phoenix2 is the occurrence; this entity corresponds closely to what is called the fiche in DEAF-DWS. Moreover, we identified a large overlap of metadata that are associated to DocLing occurrences and DEAF fiches, respectively. These metadata can easily be mapped onto each other: the written representation of the lexeme (called surface in Phoenix2, Zettelwort in DEAF-DWS), the part-of-speech information, the siglum (the abbreviation used for the text), the text-reference, the date of the text, and the scripta (i.e., the written form of a spoken dialect of Old French). Also, every DocLing occurrence is assigned to a lemma, and so is the DEAF fiche. The fiche is the basic data unit of the DEAF. It is the starting point for the editorial process. With the data import, the DEAF-DWS turns a DocLing occurrence into a DocLing fiche to make it compliant with the DEAF fiche. Conceptually, the DEAF-DWS treats DocLing fiches as a sibling type to the original DEAF fiche (i.e., they share a common supertype fiche). Each DocLing fiche corresponds to exactly one occurrence in DocLing. To allow for a mapping of a given DocLing fiche to the corresponding occurrence during the write-back to Phoenix2, each DocLing fiche contains the occurrenceID of its associated occurrence as an attribute. The occurrenceID is a unique identifier for each DocLing occurrence within Phoenix2 that had been imported.

The only integration-relevant difference between the data models, prior to the integration, was the handling of the lemma structure. DEAF-DWS models the lexemes as word families with one main-lemma and one to many sub-lemmata. In contrast, Phoenix2 categorized according to sub-lemmata, only.
To allow for a unique mapping of lemma entities in Phoenix2 to lemma entities in DEAF-DWS, we added the main-lemma to the Phoenix2 data model too.

The prerequisite for the data exchange is a compliant set of lemmata for both DocLing and DEAF. Originally, the DocLing data were lemmatized according to the Modern French orthographical norm. Thus, a preparatory step for the integration was to migrate the Modern French lemmata to Old French lemmata. This manual re-lemmatization necessarily followed the standard lemmatization of Old French conducted by the DEAF. This lemmatization is widely accepted as the norm of middle 12th century French. During this preparatory step, the 800,000 occurrences in Phoenix2 were attached to approximately 5,300 Old French lemmata.

4.2 Import and Export of Data

The applications communicate via a REpresentational State Transfer / REST web service (cf. Fielding 2000). The compliant set of lemmata of DocLing and DEAF is the foundation for the data exchange: For each lemma, we import the attached occurrences from Phoenix2 into DEAF-DWS. This import includes the following information that is assigned to a given occurrence in Phoenix2: surface, lemmaPOS (part-of-speech), sigel (siglum), division (text-reference), date, and scripta. These information units correlate with the DEAF data structure. In addition to this, the import includes scriptorium (where the document was written), context (the textual context of the occurrence), URL (of the electronic edition within the DocLing website), and, finally, the occurrenceID.

After its initial import into the DEAF-DWS, an occurrence ‘exists’ twice, i.e., as an occurrence in Phoenix2 and as a fiche in DEAF-DWS. To prevent data inconsistencies between these two representations of the same occurrence, each metadata property of an occurrence can only be modified by exactly one of the two systems. More specifically, only the DEAF-DWS is allowed to change the lemma assignment of an occurrence (technically, of a fiche that corresponds to an occurrence). All other metadata except the lemma must only be changed by Phoenix2, e.g., the date, scripta and scriptorium. After every data modification, the systems need to be synchronized in order to make the modification visible to both systems.

4.3 Workflow and Graphical User Interfaces

The workflow comprises (i) the import (from Phoenix2 into DEAF-DWS), the assignment and write back (to Phoenix2) of DocLing occurrences, and (ii) the graphical and semantic integration of the data into the respective dictionary entries. It consists of automated and manual steps. We have implemented a number of features including the necessary GUIs to the DEAF-DWS to perform these steps.

4.3.1 Lemma Assignment

The import of DocLing occurrences into the DEAF-DWS is triggered by hand. Figure 1 shows the feature implemented for the import: The editor searches for a given lemma (in the respective field, e.g., dame) and imports all occurrences attached to the lemma. Technically, the lemma assignment is not carried out on the level of the lemmata, but on the level of the occurrences. However, the GUI displays the lemmata as the unit that is most familiar to the editor. Below the search field, the GUI displays the pending lemmata, i.e., DocLing lemmata that have not been assigned to a DEAF lemma (Fig. 1).
Figure 1: Import of DocLing occurrences (via a lemma).

By clicking on the button for the lemma assignment (‘manuell zuordnen’), the interface as shown in Fig. 2 opens up. The assignment needs to be done manually whenever the given lemma is imported for the first time. The DEAF system supports this step by suggesting a main-lemma-sub-lemma combination it finds within the DEAF data where the sub-lemma matches the incoming DocLing lemma. In our example in Fig. 2, the lemma to assign is succession (subst. fem.) that is a sub-lemma of the main-lemma succeder (verb) within the DEAF-DWS.

Figure 2: Assignment of DocLing occurrences to a DEAF main-lemma-sub-lemma combination.
In cases when the system finds several homonyms for the given lemma it will display all possibly fitting main-lemma-sub-lemma combinations for the editor to choose from. In cases when none of the suggested combinations is the correct one or the lemma in question is yet unknown to the DEAF system the editor can manually create a main-lemma-sub-lemma combination (Fig. 3).

The result of the lemma assignment is presented in a second table (not shown). The export of the data to Phoenix2 is again triggered by hand (it can also be done at a later date). Via the REST web service, the export writes back the main-lemma-sub-lemma combination for each DocLing fiche to the corresponding DocLing occurrence in Phoenix2. The lemma assignment needs to be performed only for the first import. When a second import of the same lemma is triggered, the DEAF system will perform the assignment automatically based on the already existing information in DEAF-DB. As mentioned earlier, after the successful import-export procedure each Phoenix2 occurrence now has a corresponding DocLing fiche representation in DEAF-DWS.

To be able to track all imports and write-backs, detect errors, etc., all processes are stored in a third table (not shown).

4.3.2 Integration into a Dictionary Entry

After the data import and lemma assignment, the corpus data need to be merged into the respective dictionary entries. For this purpose, the DEAF-DWS displays the DocLing fiches within two GUIs, one for the editing of the graphical apparatus and one for the editing of the semantic part of the dictionary entry.
Figure 4: Editing feature for the graphical apparatus of a dictionary entry.

Figure 4 shows the GUI for the editing of the apparatus of graphical variants. The DEAF-DWS classifies the graphical realizations attested by the imported DocLing data within the apparatus. Using *surface* of the DocLing fiche and *Zettelwort* of the DEAF fiche, it arranges all graphical variants in alphabetical order and merges the DocLing data with the original DEAF data. Within the alphabetical order it collates the attestations in chronological order and, as a third assorting step, also in alphabetical order of the sigla (using the DEAF metadata units *Datierung* and *Sigel*/DocLing metadata units *date* and *sigel*). This is done in a fully automated way. Moreover, options for a manual post-processing are also provided. Note that the merged DEAF fiches and DocLing fiches are displayed within the same table for the convenience of the editor. To be distinguishable, DEAF fiches and DocLing fiches are displayed in white/blue and in shades of green, respectively.

The semantic integration of the DocLing material needs to be performed manually with linguistic expertise, and this process has recently been started. The editing process of the entries of DEAF*pré*, on the other hand, were completed in 2017. As such, the starting point for the semantic integration of the DocLing data into an entry is a completed DEAF*pré* entry with the semantic tree already established. Figure 5 shows the semantic tree (‘Bedeutungsbaum’) for the lexeme *succession* (subst. fem.) with one main-sense and one sub-sense. We can see that the (white and blue) DEAF fiches have already been assigned to the proper sense and they are displayed in a table on the right-hand side of the GUI. The newly added (green) DocLing fiches initially appear in the table on the left-hand side. Each of them needs to be assigned to the correct sense of the semantic tree. The GUI offers several drag-and-drop mechanisms and other features to do this in a time-saving way.

As soon as all DocLing fiches are merged, the updated dictionary entry can be exported as an XHTML file that is used for the online publication.
5 Online Publication

With the integration of the DocLing material, the entries of DEAFpré are enriched with attestations that are integrated into the apparatus of graphical variants and into the semantic part.

As we have shown above, the graphical integration of the corpus material is performed automatically by the DEAF-DWS. However, the manual integration of the DocLing material into the semantic structure of each article is time-consuming, and will be performed gradually in the years to come. To compensate for the long-lasting workflow, we implemented two release steps for the publication. As a first release step, we created a preliminary publication that is the result of a fully automated process. This enables us to give online access to the valuable new material before the task of the semantic integration will be accomplished. We execute the second release step after the manual post-processing has been completed, i.e., after the semantic integration of the DocLing data.

5.1 Release Step #1: Automated Processing

For the display of the new material, we have modified the design of the online publication of DEAFpré. The modification of the entry’s graphical apparatus was straightforward. We display each DocLing attestation with its siglum and text-reference as we do with the original DEAF attestations (Fig. 6).

Figure 6: DEAFpré entry mouture (subst. fem.): (headword, dictionaries, secondary literature, and) part of the graphical apparatus.
To enable the user to recognize the origin of the DocLing material we display it in a color different from the one used for the DEAF attestation. This is important because the DocLing material is of a significantly better quality compared to the original DEAFpré material. The attestations given in a DEAFpré article are not verified in the sources, i.e., in the editions of the primary texts. The reason for this major flaw is the very limited time that the two-fold concept described above allowed for the editing of the DEAFpré articles (and it is the most significant difference to DEAFplus where every information is verified). In contrast to this, the DocLing material is sound evidence that deserves to be identifiable as such.

The fact that we import the context of each attestation, the URL and the other metadata from Phoenix2 allows us to make this information accessible to the user. By clicking on any DocLing attestation, the user can display this information (Fig. 7). The button ‘Ouvrir ce passage dans DocLing’ provides the hyperlink to the respective document within the DocLing website.

For the creation of the semantic part of the entry, the export routines of the DEAF-DWS place all DocLing material (again with attestations and text-references) that is not yet properly semantically integrated into a container. In the online publication, we display this container as clearly distinguishable addenda (‘Identificanda DocLing’) to the semantic tree of the respective entry (Fig. 8).

Figure 7: Display of a DocLing fiche within DEAFpré.

Figure 8: DEAFpré entry mouture (subst. fem.): semantic part with container ‘Identificanda DocLing’.
5.2 Release Step #2: Manual Post-Processing

The second release step results from the accomplished task of the semantic integration (and thus it does not concern the graphical apparatus). The publication merges all attestations in the semantic tree, as shown for the first main-sense “travail de moudre du blé et sim.” of the lexeme *mouture* in Fig. 9.

Figure 9: DEAF pré entry *mouture* (subst. fem.): semantic part with semantically integrated DocLing attestations.

6 Added Value for Both the Dictionary and Corpus

Our aim is to merge the data with mutual benefit for the corpus project and for the dictionary. From the dictionary’s point of view, the vocabulary of the DocLing corpus texts enriches the dictionary’s material with the medieval language of law which previously had been widely unregarded by Old French lexicography. It helps to develop the comprehension of the lexemes in a considerable way, producing added value within the historical lexicography of Old French. This clearly extends the limits of the traditional historical dictionary. The added value is also of specific interest for the historical sciences focusing on medieval law and the application of law that is witnessed in the documentary sources. We foresee that the new source material will shed light on the senses of many lexemes of DEAF pré, in particular because it represents the discourse tradition of legal documents. Therefore, the DocLing material will add to a better understanding of the semantic scope of these lexemes. As a consequence, the editor’s task while integrating the DocLing material will be to evaluate the semantic tree of the DEAF pré entry and to improve and expand it if necessary. This will increase the quality of the DEAF pré entries in a significant way.

From the corpus’ point of view, one benefit is the semantic disambiguation of the corpus data. Traditionally, digital text editions – both as a single publication and as a part of a larger corpus – are stand-alone products. The publication usually does not offer an instrument (e.g., a comprehensive glossary) that supports the reader to understand the text. With the integration of the data into the dictionary we create a means that helps the reader to grasp the comprehension of the vocabulary. Also, with the integration, we embed the specific juridical vocabulary of the DocLing texts within the overall system of the Medieval French language as it is established by the DEAF. This reveals the place of the lexical units attested in the corpus within the broader semantic range of the Old French lexicon and the significance of the vocabulary within the history of the language.

7 Establishing Links from the Corpus to the Dictionary

A remaining issue is how to establish a link from a given word occurrence within a DocLing text edition (Fig. 10) back to the publication of the respective entry in DEAF él.
Based on the successful integration of the DocLing data into the DEAF database and the semantic mapping of the lexical units, we foresee the possibility to establish an automated interlinking process using the occurrenceID of each DocLing fiche and DocLing occurrence, respectively. This needs to be further evaluated in a follow-up study.

Independent from the above-described data integration is an approach that parts from the XML/TEI data of a digital text edition, as described in Tittel, Bermúdez-Sabel and Chiarcos (accepted paper): With the insertion of RDFa compliant attributes (cf. Herman et al. 2015) into the existing XML elements, the data of the text edition can automatically be enriched with hyperlinks to the DEAF dictionary. The fact that the DocLing corpus texts are published in an XML/TEI format (TEI Consortium 2017) makes this a promising approach for the creation of references from DocLing to DEAF.

8 Conclusion

To the best of our knowledge, this is the only example of a retrospective and successful integration of two voluminous and longstanding projects of a historical (Romance) language both from a graphical and a semantic point of view. We have implemented a semi-automated process resulting in a time-saving editorial workflow. We argue that our approach to fully integrate the corpus data of DocLing is a promising way to solve the problem of semantic mapping. We show how to perform the semantic mapping of the lexical units of DocLing using an existing dictionary writing system. A flexible publication concept with two release steps compensates for the time-consuming semantic integration, as it makes it possible to publish two versions of each dictionary entry: one that is created in a completely automated way, and another that shows the result of the manual post-processing with linguistic expertise. This clearly contradicts Asmussen 2013: 1087: “Combining existing dictionaries with existing corpora will inevitably yield products of second quality”.

Figure 10: Online edition of a document in DocLing.
Also, we conclude that the integration of DocLing and DEAF is a promising pilot project for the integration of other corpus linguistic data into a dictionary. At the same time, it emphasizes the role of the DEAF as a standard reference that can also be used for other single scholarly editions of Old French texts that are digitally published.

References


