

Preserving Semantic Information from Old Dictionaries: Linking Senses of the *Altfranzösisches Wörterbuch* to WordNet

Achim Stein

Institute of Linguistics
University of Stuttgart (Germany)
achim.stein@ling.uni-stuttgart.de

Abstract

Historical dictionaries of the pre-digital period are important resources for the study of older languages. On the example of the *Altfranzösisches Wörterbuch*, an Old French dictionary published from 1925 onwards, this contribution shows how the printed dictionaries can be turned into a more easily accessible and more sustainable lexical database, even though a full-text retro-conversion is too costly. Over 57,000 German sense definitions were identified in uncorrected OCR output. For verbs and nouns, 34,000 senses of more than 20,000 lemmas were matched with GermaNet, a semantic network for German, and, in a second step, linked to synsets of the English WordNet. These results are relevant for the automatic processing of Old French, for the annotation and exploitation of Old French text corpora, and for the philological study of Old French in general.

Keywords: sense extraction, sense distinction, dictionary, Old French, WordNet

1. Introduction

1.1. Historical lexicographic resources and sustainability

Historical dictionaries are a valuable resource for historical and diachronic linguistics. But the era of big lexicographical projects is definitely over. Many of the dictionaries that cover earlier periods of individual languages were compiled in the 19th and the first half of the 20th century. Whereas modern dictionaries are gradually developing towards online data bases, historical dictionaries that were compiled in pre-digital times haven't followed. Digitalizing them would be costly, and even if funds were available, many of them would require updating before doing so.

The methods described here contribute to the preservation of these lexical resources. Taken separately, they are not new, but considered together, they show how even a fully automated process can substantially enhance the usability of older lexical resources.

1.2. The case in point: *Altfranzösisches Wörterbuch* by Tobler et al.

The *Altfranzösisches Wörterbuch* (Tobler and Lommatzsch, 1925ff), henceforth “TL”, is one of these dictionaries. The project was started as early as 1857 and completed only in 2002. The first printed volumes are not available any more. For the study of Old French (OF), i.e., the language of French texts in the period between the late 9th and the early 14th century, the dictionary is extremely valuable. It is the biggest lexicographic resource for OF, with more than 52.000 entries in 11 volumes, and a large number of quotations for each sense.

In order to compensate for the unavailability of the printed first volumes, a digital version was published on a DVD containing JPEG images of the scanned pages, but not the full text of the dictionary articles (Blumenthal and Stein, 2002).¹ This digitalization method is less costly than full-

text digitalization that relies either on manual correction of OCR output or on manual copying, as for example the “double-keying” method (Haaf et al., 2013) (for a feasibility study on the retroconversion of a big French etymological dictionary see (Renders, 2015)).

The data of the TL DVD edition were the input for the sense extraction procedure described here.

1.3. The Challenge

From a philological perspective, the TL is one of the two most important dictionary resources for OF, besides the Godefroy (Godefroy, 1880ff). Its most valuable contribution to the research on OF are the fine-grained sense distinctions paired with rich quotations for each sense. However, the word senses are given in German, which limits the international usability of the dictionary considerably. On top of this, the vocabulary used in the German translations is often obsolete, and sometimes hard to understand even for German speakers, especially in the earlier volumes.

From a technical perspective, the most obvious issue is the lack of full-text access. The DVD provides access to the image files via a lemma list that was extracted manually, but the creation of a full-text version has never been undertaken by the publisher, for several reasons. One is the typographical quality, which varies considerably between the eleven volumes and makes text conversion using optical character recognition (OCR) error prone. On the other hand, most researchers working on OF content themselves with the relatively comfortable access to the page images. But the lack of a full-text version makes it difficult to extract the semantic information from the resource in a reliable way.

1.4. Solutions

This contribution reflects our opinion that dictionaries of this kind need to be transformed as well as possible into sustainable lexical resources, and even more so if it is unre-

¹The DVD can be purchased from the publisher, Steiner Verlag, Stuttgart, Germany (link in the bibliography). It contains all

but the last volume of the TL. The missing articles were replaced by their equivalents from Godefroy's dictionary.

alistic to create a fully fledged digital version. For the TL, we pursue the following objectives:

- modernization, by converting and updating the meta-linguistic information;
- sustainability, by linking the sense descriptions to a standardized, widely accepted conceptual hierarchy: GermaNet (the German WordNet);
- internationalization, by providing links to the Princeton English WordNet.

Ideally, the result is more than an update of the information contained in the dictionary. It should rather provide us with an onomasiological access based on a conceptual system, and thus allow us to retrieve synonyms, similar meanings, hyponyms, or semantic classes. The added value for research on OF is considerable.

Section 2. explains how we analysed the resource and extracted the sense descriptions for each lemma, section 3. discusses the problems and the results, and section 4. concludes.

The procedures described here are, with few exceptions, fully automatic and unsupervised. They can't be expected to produce a fully reliable digital resource. In order to allow the community to improve the resource, we made all the data available that are needed to reproduce the steps described here (see section 6.).

2. Sense Extraction

2.1. Procedure

Sense extraction was performed on the uncorrected text output produced by optical character recognition (OCR).² Manual correction was restricted to the cases where the software had failed to recognize the two-column format. All other modifications were done automatically and can be reproduced using the script and the OCR output files.

2.2. Identifying Lemmas (Articles)

The HTML format produced by the OCR software was simplified so that only the markup relevant for the structure of the dictionary was preserved. The script reads the list of lemmas distributed with the DVD version in order to identify the position of each lemma in the text files. It compares each lemma from the list with the (supposed) lemma string in the dictionary file, using the Levenshtein algorithm in order to cope with eventual OCR errors. The script also introduces further markup containing the lemma positions and other information like the degree of Levenshtein similarity.

²The best results were achieved using *Abbyy FineReader* v12.1.6 on the original high-resolution TIFF image files (the published DVD version contains compressed JPEG images in lower quality). The resulting uncorrected text files were made publicly available by permission of the publisher (Steiner Verlag, Stuttgart, <http://www.steiner-verlag.de/>). The export format was set to HTML in order to mark font changes, while preserving the original page, line and word breaks.

2.3. Identifying Senses

The major difficulty concerning the retrieval of senses comes from the use of italics as typographical markup in the print version. First, italics are not limited to word senses, i.e., the German translations, but they also mark grammatical information (e.g., *transitive*), contextual information (typical arguments), and information on domains, registers and the like. Second, italics are a notorious problem for OCR: the error rate is considerably higher than with 'normal' font. Thus, although the OCR software was quite good at distinguishing italic from normal font, the errors and the internal structure of the passages impeded the direct processing of the senses (translations).

Extraction of the potential senses was performed for all lemmas, but their further treatment is limited to verbs, nouns, and adjectives, i.e., the categories that are covered by GermaNet. Special attention was given to the senses of the over 7000 verbs in the TL, first, because verbs are more polysemous than other categories, and second, because we hope to be able to enrich the information on verbs in a way that will benefit the analysis of other OF resources, e.g., by adding the information on arguments provided in GermaNet, in WordNet, or in other resources that can be linked to these semantic networks.

The identification and analysis of the typographically marked strings is crucial for the success or failure of the sense extraction. At the time of writing, the script only corrects some obvious and frequent OCR errors. Meta-linguistic information (*trans.*, *intr.*, *fig.* etc.) is removed. The reflexive pronoun (*sich*) is ignored, since it is not present in GermaNet verb entries.

Parsing of the potential senses is quite rudimentary. We apply two simple rules: First, a comma within the string is taken to separate two German translations of the same sense, regardless of the category. Splitting the string into sub-senses using commas mostly yields correct results, except when enumerations occur within a sense. Second, if a sub-sense has no direct match in GermaNet (which is normally the case when it consists of more than one word), a simple, category-sensitive rule checks if the last word of verbal senses or the first capitalized word of nominal senses has a GermaNet match (excluding classifier nouns like *Art*, as in *eine Art von ...* 'a kind of ...', and support verbs like *bringen*, *sein*, *werden*, etc.). This second rule takes advantage of the fact that the senses are in German: nouns are normally capitalized, and in an OV language the verbal infinitive normally occurs in final position, as in *Laute_N spielen_V* 'play the lute'. The application of these rules is marked by a code in order to facilitate the evaluation of their success.

2.4. Linking Senses to a Semantic Network

Since TL senses are given in German, GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) is the obvious resource to provide a structured concept-based access to the dictionary. GermaNet is the semantic resource with the widest coverage for verbs, nouns, and adjectives. Version 14 has 14,178 synsets with 18,336 lexical units for verbs, 108,323 synsets (139,397 lexical units) for nouns, and 13,762 synsets (17,267 lexical units) for adjectives.

GermaNet has been integrated into EuroWordNet, a multi-lingual lexical database that is linked to the original English Princeton WordNet (Fellbaum, 1999) via interlingual indexes. A research license free of charge is available from the University of Tübingen.

The joint use of GermaNet and Wordnet helps attain the goals presented in section 1.4. as follows:

1. Modernization, by linking the strings that are potential TL senses to a GermaNet entry. This goal can fail (a) when no GermaNet entry exists, (b) when the lemma in the string is not identified, or (c) when the string is not a sense (but some other piece of information printed in italics).
2. Internationalization, by finding a corresponding WordNet synset via the German-English interlingual index. This goal will fail if interlingual index doesn't have an entry for the GermaNet lexical unit.

Both of these procedures are executed in the respective subfunctions of the script. They require the following resources (all freely available): the TL text files including HTML markup as described in section 2., the GermaNet XML files, the WordNet 3.0 index and data files, and the interlingual index (part of the GermaNet distribution). The problems encountered with these procedures are described in detail in the following subsections.

2.4.1. Modernization: Linking Senses to GermaNet

We briefly present typical cases that cause the identification of a GermaNet verb entry to fail.

(a.) Collocations and fixed expressions. In some cases, the expression is merely a specification of the verb sense, as in the entry *tabloier*: for the sense *ein Brettspiel spielen* 'to play a board game' the verb *spielen* alone is still an acceptable GermaNet equivalent. In the more opaque cases, like *die Trommel rühren* (lit. 'stir the drum'), the verb *stir* taken alone yields an incorrect sense. Often, as in this example, the use of the verb in these contexts is obsolete (instead of Modern German *die Trommel schlagen*).

(b.) Domain classifiers. Some 'senses' only indicate the domain in which the sense is used, e.g., for OF *taborer* the indication *Falkenjagd* 'falcon hunt' or for OF *taire* the domain *Rechtssprache* 'legal language'.

(c.) Meta-linguistic information. Comments on usage, e.g., *als Eigennamen* 'as proper name', remarks on other dictionaries (e.g., *godefroy vii 661a unrichtig* 'Godefroy... incorrect').

More than 2,000 obsolete translations or fixed expressions were substituted with a Modern German equivalent (without modifying the original data) in order to match a GermaNet entry. Apart from the column recognition errors mentioned in section 2.1., this was the only manual correction applied to the OCR output.

2.4.2. Internationalization: Linking GermaNet to WordNet

In GermaNet, synsets have an ID (e.g., "s58444"), a category (e.g., "verben"), and a class (e.g., "Perception"). Lexical units have an ID, a number for one or more senses, and an orthographical form. The interlingual index relates the

ID of the lexical unit to a word and an ID in the Princeton WordNet 3.0 via a semantic EuroWordNet relation (e.g., "synonym", "hypernym" etc.).

For example, one of the senses of the TL entry *abaissier* (cf. ModF *abaisser*) is translated by German *unterwerfen*, which is listed in GermaNet as a 'competition' verb. In the interlingual index, the ID of the GermaNet lexical unit ('179915') is linked to the WordNet 3.0 synset '02424128-v' by a 'synonym' relation. The target synset in WordNet subsumes a set of verbs (*keep down, quash, reduce, repress, subdue, subjugate*). The goal of internationalizing the access to the dictionary is reached, since any of these English verbs can serve to retrieve this sense of OF *abaissier*.

However, successful internationalization is restricted by the coverage of the interlingual index. GermaNet (v14) has 18,335 lexical units for verbs. The interlingual index has 28,565 records, but only 4,986 refer to a verbal synset in WordNet (v3.0). So we can expect to reach the goal "internationalization" only for a subset of the GermaNet entries. There are other ways to link the German synset to a WordNet synset, e.g., by testing synonyms of the German verb, or by establishing the link for a hypernym, but the risk to identify a WordNet concept that is too general or semantically inappropriate would increase, so we won't pursue this option further here.

The result, i.e., the GermaNet or WordNet synset and the semantic relation, is inserted into the markup of the TL text files. In order to facilitate the evaluation of our results, the list of verb senses and the corresponding GermaNet/WordNet synsets was uploaded in spreadsheet format (see section 6.). The discussion of the results in the following section is based on these data.

3. Discussion of Results

3.1. Quantitative Evaluation

Table 1 lists the results per sense. The first column distinguishes the part of speech categories (verb, noun, adjective). The second column gives the number of "senses", i.e., strings that were identified as potential senses according to the procedure described in section 2.3.. Columns 3 and 4 give the absolute number and the percentage of senses that were successfully linked to a synset in GermaNet. Column 5 gives the absolute number of senses that were successfully linked to WordNet. Column 6 gives the percentage of GermaNet matches that were linked to WordNet synsets. Note that adjectives are not present in the interlingual index, which explains why only three synsets could be linked to WordNet (they were listed in entries for nouns).

pos	senses	GermaNet		WordNet	
verb	18363	13936	75.89%	5803	41.64%
noun	31856	20777	65.22%	12081	58.15%
adj.	7430	2680	36.07%	3	0.11%

Table 1: Evaluation per sense: matches in GermaNet and WordNet

Table 2 aggregates per lemma. The first column distinguishes the part of speech categories (verb, noun, adjective). The second column gives the number of lemmas per

category. Columns 3 and 4 give the number and percentage of lemmas without GermaNet match, i.e., when none of the senses identified for this lemma could be matched with a GermaNet synset.

pos	lemmas	no GermaNet match	
verb	6759	449	6.64%
noun	17521	3426	19.55%
adj.	3329	1507	45.27%
total	27609	5382	19.49%
total	28,028	6,197	22.11%

Table 2: Evaluation per lemma: percentage of lemmas without a GermaNet match

Considering that (a) OCR output was not corrected, (b) sense identification was based only on the recognition of font changes, (c) no particular effort was made to apply parsing or other NLP techniques to the potential senses, (d) some of the extracted “senses” are not real senses, the score of automatic sense attachment to GermaNet is quite satisfactory. The suspected higher complexity of verbal senses, as compared to nominal senses, does not seem to affect the attachment score. However, nouns fare much better than verbs in the next step, i.e., the link from GermaNet to WordNet. This is due to the better coverage of nouns in the interlingual index.

Verbs, on the other hand, are normally more polysemous than nouns, and this is why the aggregated results per lemma (Table 2) are much better for verbs: only for 6.6% of the verbs none of the senses could be linked to GermaNet, whereas almost 20% of the nouns remain without link. We take the per-lemma results to be particularly relevant, because for many applications it will suffice to attach one or two senses to a synset in order to achieve a semantic classification (retrieval of semantic verb classes, corpus annotation using semantic class labels, etc.). Manual verification will be needed in order to check if the subset of senses per lemma that has a GermaNet link are also the most indicative ones for their core meaning.

3.2. Qualitative Evaluation

A manual evaluation of the plus 57,000 senses that our script has identified across all categories was not feasible. In this section, I first discuss some typical scenarios which hampered successful sense extraction (3.2.1.) before presenting the results of a sample evaluation (3.2.2.). In both cases, I will focus on verbs. With verbs, it is more difficult to identify the correct sense, first, because they are generally more polysemous, second, because verbal polysemy also affects the argument structure, which is not only captured lexically (i.e., by the choice of the German equivalent in the definition), but also by additional grammatical markers (‘reflexive’, ‘transitive’, etc.) or complex verb constructions (e.g., combinations with German *lassen* ‘let’ to mark causative constructions).

Data and examples discussed below are taken from the results achieved at the time of submission and can be verified in the published list of extracted senses (cf. section 6.).

3.2.1. Typical problems or failures

The following success and failure scenarios will facilitate the understanding of the procedures described above, and allow me to point out some possible future improvements.

a. Fully Successful Retrieval. For *baillir* ‘administer, care for’, both senses were retrieved and matched with a GermaNet synset and linked to a WordNet synset. The first sense *verwalten, regieren* was linked to English *administer* and *govern* via synonym links (which are preferred over other links in case of multiple matches). Both of them are good sense equivalents. The second sense, *behandeln, bedenken*, was linked to English *care for* and *consider*. *Consider* is a good fit, but *care for* is not. It was chosen because polysemy within GermaNet is not yet properly accounted for: *behandeln* ‘care for’ is listed as a social sense and was matched first, whereas the cognition sense ‘deliberate, go about’ was only the second match. The hierarchy of WordNet synsets offers obvious ways of identifying possible interlingual mismatches of this type and implementing preference rules: *consider* and *deliberate*, for example, are subsumed under the same synset (00814706 in version 3.1) whereas *care for* is hierarchically more distant.

b. Partly Successful Retrieval. An example for partial success is the verb entry *bäaillier intr. gähnen*, with a one-verb German sense definition following a meta-linguistic abbreviation (here: intransitive). The German verb is listed in GermaNet and linked in the interlingual index to *yawn* via a synonym link to the WordNet synset. The TL gives three further senses of *bäaillier*. They are specifications of the first sense and defined by adverbial prepositional phrases like *vor Hunger* ‘of hunger’. In this case the GermaNet match would suffice the purpose of sense or verb class identification. In other cases, like *im Todeskampf* ‘in death struggle’ it probably wouldn’t because *yawn*, just like Modern French *bâiller*, is not a contextually equivalent translation. Anyway, the lack of a verb in the definition prevents us from linking it to a more fitting concept like *gape*.

c. Failed Retrieval. Most of the lemmas for which none of the senses could be matched with GermaNet (cf. Table 2) have only one or few senses. Sometimes the reasons for failure are trivial: For *asailir*, the first sense (‘attack’) was not retrieved, because the font change was not properly recognized or encoded during OCR. The second sense consisted only of a grammatical class (*intr.*), without a verbal equivalent.

Even without the analysis of a significant sample, an eye-ball estimation shows that by far the most frequent reason for failure are translations with a predicate consisting of an adjective and a support verb, as in *asserir still, einsam werden* ‘become silent, lonely’. They are most common with verbs expressing a state or an (inchoative or causative) change of state, and only for a subset of them a lexicalised (mostly prefixed) verb is available (e.g. for *dunkel werden* ‘become dark’ > *verdunkeln* ‘darken’, but not for *still werden*). The non-availability of a simple verb prevented successful matching with GermaNet, since support verbs like *werden, sein, machen* were ruled out as potential matches for good reason (cf. section 2.3.).

3.2.2. Manual validation of 1000 verb senses

For the 417 verbs beginning with letter 'C' (TL volume 2), 1,055 (potential) senses were identified. For 73 senses, no GermaNet match was found. For senses with more than one German translation (cf. section 2.3.), each suggested GermaNet match was validated, yielding a total of 1265 matches. We³ validated the quality of the suggestions using a validation scale from 1 (best) to 5 (worst), the values indicating whether the GermaNet match is

1: correct (a semantic and grammatical equivalent);

These are basically verbs without any potential grammatical ambiguity (as opposed to value 2), like *conforter* 'ermahnen' ('admonish').

2: lexically correct, but grammatically divergent: the argument structure of the match was different or unspecified;

The vast majority of these cases are reflexive meanings. Since GermaNet has no special entries for reflexive verbs (with the pronoun *sich*), these senses are normally matched with the corresponding transitive verb sense. For example, for OF *contraitir*, the meaning 'sich zusammenziehen' (unaccusative, 'something contracts') is matched with the transitive (causative) GermaNet meaning 'to contract something'.

3: a close hypernym of the sense;

Example: OF *convertir* 'convert' has a sub-sense 'sich (in christlichem Sinne) bekehren' ('to become converted in the Christian sense').

4: semantically similar, but not a close hypernym;

These are mostly very specific verb meanings. For example, OF *croire* 'believe' has a specific meaning 'sich jem. verpflichtet erkennen' ('to realize one's obligation to sb.'). Here, the default verb sense 'believe' can't be considered as a close hypernym.

5: incorrect (too distant or unrelated).

Example 1: For OF *cometre*, the German translation 'begehen' is ambiguous between 'commit' and 'walk along'. The latter was (wrongly) selected.

Example 2: As mentioned in section 2.4.1., most verbs in verb-noun collocations have incorrect matches. For OF *confesser*, the German translation is 'die Beichte abnehmen' ('to hear sb's confession'). The meaning(s) of the match *abnehmen* ('take away' or 'diminish') are too distant from the meaning of the multi-word sense.

Table 3 shows the result of the sample validation. For each value (1-5), it gives the absolute number and the score relative to the total of validated GermaNet matches.

The average value of all validations is 1.6, and about two thirds of the GermaNet matches are 'correct' (value 1).

³My validation was cross-validated by Yela Schauwecker, researcher in the DFG BASICS project (cf. section 5.). Cases of inter-annotator disagreement were discussed and resolved.

value	senses	score [%]
1	879	69.5%
2	159	12.6%
3	124	9.8%
4	18	1.4%
5	85	6.7%
	1,265	100%

Table 3: Validation of GermaNet matches for verb senses (letter C)

What is more important, however, is usability. A close hypernym (value 3) or a match that only differs with respect to its construction (value 2) still allows for the correct assignment of semantic verb classes in a corpus annotation task for example. A semantically similar match (value 4) might still be usable, whereas the worst case (value 5) clearly is not. Values 4 and 5 add up to 8.1%, so the final verdict, at least for the validated sample, is satisfactory, even more so because the score of correct matches can be assumed to be higher for nouns than for verbs, for the reasons given at the beginning of this section.

4. Conclusion

On the example of an Old French dictionary published from 1925 onwards, this contribution shows how older printed dictionaries can be turned into more easily accessible and more sustainable lexical databases, even when a full-text retro-conversion is too costly. The added value is considerable, and the TL is probably the first historical language resource that was connected to a semantic network.

The method described here is based on unsupervised (and to a large extent uncorrected) OCR of the printed text followed by an automatic interpretation of the structure of the articles relying on the detected font changes. The output is interpreted linguistically, with the goal of connecting for each lemma as many senses as possible to a widely used semantic network. Since the sense definitions of the *Altfranzösisches Wörterbuch* are in German, senses were matched with GermaNet entries and, where possible, linked with synsets of the English WordNet.

The focus of this contribution lay on the lexicographic domain, i.e., on making old resources more sustainable, rather than on the domain of semantic networks. Further improvements of the automatic procedure could include: (a) a more elaborate parsing of the sense definitions and conversion of the most frequent patterns of multi-word expressions into simple lexemes that are listed in GermaNet (e.g. support verb constructions into prefixed simple verbs); (b) a better treatment of polysemous entries in GermaNet in order to retrieve all the matching WordNet synsets.

Nevertheless, even without exploiting these possibilities, the results of the sense extraction were quite satisfactory. More than 57,000 senses were extracted, and for about 93% of the verbs and 80% of the nouns, at least one sense was matched with GermaNet, thus providing, for over 20,000 lemmas and 34,000 senses a semantic access to the dictionary entries (verbs and nouns), as well as the possibility to extract senses by semantic class using the Ger-

maNet/WordNet hierarchy of synsets. The manual validation of about 1000 verb senses has shown that the dictionary senses were matched with GermaNet senses with satisfactory quality. On top of this, about half of the GermaNet entries could be linked to a WordNet synset. The result is a considerably enhanced lexical resource whose meaning representation was made more sustainable and, partly, more easily accessible for the international community. The results and the resources needed for eventually improving them were made public.

5. Acknowledgements

Some of the work presented here was funded by the Deutsche Forschungsgemeinschaft (DFG) grant no. 265711632: *Borrowing of Argument Structure in Contact Situations (BASICS)*. German verb senses were verified and partly corrected or modernized by Katharina Fezer, who also analysed the methods and benefits of integrating GermaNet and WordNet. I thank three anonymous reviewers for their insightful comments and suggestions, which helped me improve this revised version.

6. Language Resource References

The following resources are publically available at <http://tinyurl.com/tobler-lommatzsch>:

- a. the original (uncorrected) OCR output
- b. the list of lemmas
- c. the list of extracted senses and matched synsets

7. Bibliographical References

- Peter Blumenthal et al., editors. (2002). *Tobler-Lommatzsch: Altfranzösisches Wörterbuch. 4 CD-ROMs und DVD mit Begleitbuch*. Steiner, Stuttgart. <http://www.steiner-verlag.de/titel/53155.html>
- Christiane Fellbaum, editor. (1999). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA. <https://wordnet.princeton.edu>
- Godefroy, F. (1880ff). *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX. au XV. siècle*. Paris. Kraus Reprint Nendeln/Liechtenstein 1964.
- Haaf, S., Wiegand, F., and Geyken, A. (2013). Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text. *Journal of the Text Encoding Initiative*, (Issue 4), March.
- Hamp, B. and Feldweg, H. (1997). Germanet - a lexical-semantic net for german. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics: Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. <http://www.sfs.uni-tuebingen.de/GermaNet/>
- Henrich, V. and Hinrichs, E. (2010). GernEdiT - The GermaNet Editing Tool. In Nicoletta Calzolari, editor, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2228–2235, Valletta. ELRA.

- Renders, P. (2015). *L'informatisation du Französisches Etymologisches Wörterbuch: modélisation d'un discours étymologique*. Travaux de linguistique Romane. Éditions de Linguistique et de Philologie, Strasbourg.
- Tobler, A. and Lommatzsch, E. (1925ff). *Altfranzösisches Wörterbuch*. Weidmann, Berlin u.a.