On automatically disambiguating end-of-line hyphenated words in French texts

• Pöckelmann, Marcus
  Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Germany
  marcus.poeckelmann@informatik.uni-halle.de
• Ritter, Julia
  Institute of Romance Studies, Martin-Luther-University Halle-Wittenberg, Germany
  julia.ritter@student.uni-halle.de
• Gießler, André
  Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Germany
  andre.giessler@informatik.uni-halle.de

Summary

We introduce an approach for disambiguating end-of-line hyphenated words in French based on the given text and an external dictionary. Compared to the obvious way to handle hyphenation automatically our approach has reduced the number of falsely guessed spellings for a tested set of French texts by at least the factor of 1.5. In particular, the approach helps to reduce the effort of manually determining the correct spellings dramatically by focusing on the few cases where the approach is not reliable.

In printed texts usually a lot of words are separated by a hyphen at line breaks. Such a hyphenation is made if the last word is too long for the current line particularly with regards to a justified text alignment. Whereas in many cases an additional hyphen (soft hyphen) will be appended to the first part of a word, some words already contain a hyphen (hard hyphen) that can be used for the line break. During different steps of automated text processing hyphenation can be hindering if the correct spelling of a word, whether with or without hyphen, is unknown. Just think of applications in which the text shall be annotated automatically or shall receive a different typesetting. In such cases it is desirable to use a self-acting or at least a semiautomatic approach in contrast to make manual decisions for every word’s correct spelling, which can be notably time-consuming for long texts.

There is only a sparse amount of comments in the literature how to handle the problem described above, especially in French. Some publications propose to make the decisions manually [1] [2]. The documentation of the Oxford Concordance Program [3], which is a software from the 1980s, states that it “has a facility to request that hyphenated words at the ends of lines should be reconstituted” but without giving details of the realization of this feature. One trivial procedure, removing all end-of-line hyphens, is used in a paper about tokenization [4]. This paper also mentions the use of a dictionary as possibility to reduce the error rate, which is an essential part of our approach discussed later.

Simply joining the separated parts of a word by leaving the hyphen out may solve the problem for most instances in many languages, e.g. English or German. In French however a more complex approach is necessary because the hyphen is frequently used in positions other than the end of a line. Particularly this includes the building of compounds with prefixes, nouns and pronouns as well as numbers that made their way into the written language. Whereas hyphenation in French usually follows well-defined rules nowadays, these rules changed through time and had not been applied consistently. Thus a reliable rule-based approach for disambiguating end-of-line hyphenated words is unlikely.

To solve the challenge we have developed a dictionary-based technique for reversing the hyphenation for a given French text. The approach consists of three steps. First, an internal attribution is computed which determines the number of occurrences for every word of the text under consideration. Thereby only occurrences not separated by a line break are considered. Thus the text itself will become a reference for the correct spelling of a given separated word by comparing the number of occurrences of both possible spellings. The second step is a query in an external dictionary. Again both spellings of the word, whether with or without hyphen, are searched in the given dictionary. The third step merges the information of the two previous steps in order to provide a guess for the correct spelling. Both of the previous steps may have led to either no indication at all, or to an indication for exactly one spelling or to an indication rendering both spellings probable. Thereby 16 cases are possible. Our approach assumes that the spelling with hyphen is correct if the internal attribution returns only the entry for the
spelling with hyphen even if the external dictionary says differently. This keeps the consistent spelling of the author or the age of the text. The hyphen is also chosen if the dictionary only provides this spelling and simultaneously the internal attribution either has no entry or has entries for both notations. In all other but two cases the spelling without hyphen is assumed correct. The two exceptions are the cases where the internal attribution contains both spellings and the external dictionary simultaneously provides either no or both entries. In these cases the highest number of occurrences in the internal attribution is decisive. If they are equal the spelling without hyphen is chosen.

As previously mentioned the heuristic “always use the spelling without hyphen” is the obvious way to handle hyphenation in most languages. We tested our approach against this simple heuristic with respect to the number of faulty decisions. For comparison we used a book by Guillaume Raynal in four different editions which were published in 1770, 1774, 1780 and 1820 [5] and a dictionary of the ABU : la Bibliothèque Universelle [6] with more than 250,000 entries of common words for the second step. The slightest relative difference between both techniques occurred in the edition of 1820 which contains 1,339 individual hyphenations of 52,372 words and 7,198 lines. Our approach resulted in 30 wrong guesses (2.240%) instead of 45 (3.368%) made by the heuristic which is a decline by the factor of 1.5. The biggest difference appeared in the edition of 1780 with 1,063 individual hyphenations, 44,078 words and 6,290 lines. While the simple heuristic resulted in 45 faulty decisions (3.814%), our approach nearly cut the number of errors in half to 24 (2.034%). Concerning the editions of 1770 and 1774 the outcome was 6 errors (0.819%) instead of 10 (1.364%), and 14 (1.317%) instead of 23 (2.164%), which is about the same level. The effectivity of our approach becomes apparent if the four editions are considered as one text. Our approach benefits from many words with multiple occurrences in the concatenated text consisting of 155,160 words and 21,551 lines. Only 39 (1.107%) of 3,522 individual hyphenations are reversed incorrectly. In contrast the simple heuristic makes 98 faulty decisions (2.783%).

In summary our approach dominates the simple heuristic regarding the number of wrong spellings without being free of errors itself. This is important if a researcher depends on an automated disambiguation of the end-of-line hyphenated words due to the size of the text or missing expertise for deciding the correct spelling. Furthermore our approach can be helpful to considerably reduce the manual effort scholars have for checking correctness. For the tested text all but one¹ error of our approach occurred for words without any information in the internal attribution and the external dictionary. Thus a researcher can focus on these not reliable cases instead of checking every word separated by a line break. This will reduce the effort to 298 instead of 733 individual hyphenations in the edition of 1770 (40.655%), 222 instead of 1,063 in 1774 (20.884%), 197 instead of 1,180 in 1780 (16.695%), 62 instead of 1,339 in 1820 (4.630%) and 270 instead of 3,522 in the concatenated text (7.666%).

While nearly all errors of our approach occurred for words without information in both the internal attribution and the external dictionary, skipping the second step would result only in slightly increased error rates. In contrast using the external dictionary without an internal attribution would lead to apparently more errors. Both issues may be due to the historic word forms and proper names found in the text. However a large amount of entries in the internal attribution which requires that the text under consideration is relatively large seems to be the important factor for lowering the number of errors and reducing the semiautomatic effort respectively. Keeping this in mind, the approach can easily be extended by filling the internal attribution with larger corpora so that reversing the hyphenation of a relatively small text will benefit from the advantages described above. The same can be done in step two by using multiple external dictionaries.

Acknowledgments

This research was funded by the German Federal Ministry of Education and Research (BMBF) [grant number 01UG1247] as part of the project “Semi-automatische Differenzanalyse von komplexen Textvarianten” under the direction of Prof. Dr. Thomas Bremer, Prof. Dr. Paul Molitor, Dr. Jörg Ritter and Prof. Dr. Hans-Joachim Solms. Also we would like to acknowledge and thank our project collaborator Susanne Schütz.
Notes

¹ This exception is a contradictory case for the word “par-tout” in the edition of 1780 which was guessed falsely with hyphen as it was found 17 times with this spelling in the text but only without hyphen in the external dictionary. The situation is a different one if the four variants of the text are considered in all as the spelling without hyphen was used more often in the other editions than 1780.

References

4. Gregory Grefenstette and Pasi Tapanainen (1994) "What is a word, What is a sentence? Problems of Tokenization" In third International Conference on Computational Lexicography (Complex'94):79-87,Budapest
5. Guillaume Thomas François Raynal "Histoire philosophique et politique des établissements et du commerce des Européens dans les deux Indes" - book six, in editions 1770 (Amsterdam), 1774 (The Hague), 1780 (Geneva) and 1820 (Paris)