Requirements on the Punctuation Reconstruction for the Translation of Post-modern Poetry

Meyer-Sickendiek, Burkhard

bumesi@zedat.fu-berlin.de
Freie Universität Berlin, Deutschland

Baumann, Timo

baumann@informatik.uni-hamburg.de
Universität Hamburg, Deutschland

Hussein, Hussein

hussein@zedat.fu-berlin.de
Freie Universität Berlin, Deutschland

Punctuation is an important and cohesive device in all kinds of written discourse. Standard marks used to separate words, phrases, clauses and sentences for the purpose of cohesion. Already [2][5][1] pointed out that through punctuation marks, one can signal different information structures in written language. Regarding the translation of texts, we use such marks to identify the ends of sentences, closely related sentences or clauses, etc. This is why missing punctuation burdens the translations and forces the translator to go over the text several times to understand its meaning [10]. Understanding the uses and functions of punctuation marks, therefore, is extremely important for translators, as their purpose is to clarify the meaning of a particular construction within a text. On the other hand, modern poetry often disregarded such punctuations. Ever since Italian Futurism around 1900 spoke of the 'parole in libertà', i.e. the liberation of words from grammatical and syntactic limitations, modern poetry has hardly used punctuation. This lack of punctuation makes analysis, but also translation, more difficult. The only way to reconstruct this punctuation is by listening to the poems, i.e. by subsequently identifying sentence boundaries. However, this lack of punctuation can be found very often in modern and post-modern poetry, so the challenge is to recognize the phrase boundaries. We contribute in the paper an application towards the problem of identifying left-out punctuation in post-modern poetry, by proving that only a very simple type of punctuation - the semicolon - is needed to improve machine translation. This simple punctuation refers to phrase boundaries, the so-called "grammetrical units", which Donald Wesling defined in his study "The Scissors of Meter" [11]. Such units must be identified in order to improve machine translation.

The need for adding left-out punctuation becomes in case of creating machine translations obvious with regards to the poem "bitte verlassen sie diesen raum" (english: please leave this room) written by the German poet Nicolai Kobus [6] (Text A):

bitte verlassen sie diesen raum
so wie sie ihn vorfinden möchten
danke möchten sie diesen raum
vorfinden wie sie ihn verlassen
haben bitte räumen sie alles so
vorgefundene als wären sie
verlassen worden danke sie
möchten doch nicht daß man
sie so verlassen im raum vor
findet bitte seien sie für einen so
verlassen vorgefundenen raum
dankbar [...]

The challenge for the interpretation of this poem lies in the adequate identification of the line endings. These endings can only be identified correctly by listening to the poet's reading, which is possible because we got the audio version on the *lyrikline* [7] (the world's largest corpus of spoken (post-) modern poetry which also features translations for many of the poems) webpage. This is the reason, why the manual translation, made by Catherine Hales, is able to translate these endings in a correct manner (Text B):

please leave this room
in the state in which you would like
to find it thank you would you like
to find this room in the state in which
you have left it please clear out
everything thus found as though you
had been left thank you you would not
like somebody to find you left
abandoned in the room now
would you please be grateful for
a room a space found in such
an abandoned state (...)

In the human translation or the target poem, made by Hales, there is just a little difference. This difference is caused by the missing punctuation. And it can basically be explained by the fact that Hales has chosen a different line arrangement. In terms of content, however, her translation is reproduced correctly. Since there is no specific translation system trained with poem data with/without punctuation (small amounts of training data), we used a Google machine translation (GMT) system [3]. When we compare this (human) translation with the GMT system, we recognize the difficulty of recognizing the sentence boundaries within the poem without punctuation (Text C):

please leave this room
as they would like to find him
Thank you for wanting this room
find out how to leave him

please have everything clear
found as if they were
Thank you
you do not want that one
So leave them in the room
please find one for you
leave found space (...)

Obviously, this machine translation (MT) becomes much better if we add the full punctuation marks to the source text, when listening to the audio of the poem (Text D):

please leave this room
as you would like him to find
Thank you. Do you want this room
find how they leave him
to have? Please clear everything up
found as if they were
been left. thank you
Do not want that one
So leave them in the room
please, please be for one
leave found space
grateful. (...)

Punctuation is an essential aspect of poetry translations, as it is for discourse analysis in general [8]. Punctuation “gives a semantic indication of the relationship between sentences and clauses, which may vary according to languages”, as well as to translations [4].

A first step towards solving the problem of translation unpunctuated texts is the correct localization of the missing punctuation within such sentences and clauses. In the Google translation, which was completely without punctuation, we see that Google system translated every single line anew (Text C), ignoring the line-arrangement and the “enjambments”, when one phrase continues beyond the line, or continues from the previous line. This explains the translation error in the third line: Reading the line as a full sentence disregarding its character as an enjambment, the translation produces a full sentence (Thank you for wanting this room), which does not fit to the original (... danke. möchten sie diesen raum ...). However, this translation error will be improved if we add the missing punctuation to the machine translation, which could be identified as Text D.

It is hard to translate automatically without having information about the sentence boundaries and the punctuation as a discourse unit for meaning demarcation. But to what extent punctuation information has to be recovered for the translation of post-modern poetry? Which kind of information do we need to improve machine translation? Do the questions have to be distinguished from the statements? Or is the simple marking of phrase boundaries already sufficient? To answer these questions, we analysed unpunctuated German poems. There are 234 german-speaking poets on the *lyrikline* webpage reading a total of 2591 poems. A total of 733 German poems are translated to English which are used in this work. There are 98 German poems which do not contain any punctuation information. We analysed 120 poems in this work with a maximal punctuation information ratio of 0.05%. This process yields a total of 2924 lines out of which only 28 (0.009%) with punctuation information.

The philological scholar of our project annotated the punctuation information manually by using text and audio information in the 120 poems, focusing on the intonation of poets reading their poems. In order to clarify the question which type of punctuation has to be added, we inserted two kinds of punctuation in the source text. In a first step, we focused on six different punctuation marks: full stop (.), comma (,), semicolon (;), colon (:), exclamation mark (!), and question mark (?). In a second step, we simplified this insertion by reducing these six marks to a single semicolon.

The human reference translations are compared with the automatic translation of GMT system without/with consideration of punctuation information. The experiment consists of three tasks based on the GMT system:

- Task 1: Standard translations of original poems (without punctuation).
- Task 2: Translations with one level of punctuation information: replacement of all manually annotated punctuation information by one level of punctuation (;).
- Task 3: Translations with six punctuation information: consideration of the six manually annotated punctuation information (.,;:!?).

The translation enhancement should be observable from improved translation quality scores. The results are calculated by bilingual evaluation understudy (BLEU) [9] score, which used for evaluating the quality of text by translation. The BLEU score of tasks 1, 2, and 3 are 0.256, 0.275, and 0.280, respectively. The results indicate that we need just one type of punctuation - semicolon - to improve the scoring for automatic translations of post-modern poetry.

Every generic translation system is trained with data in which segments are defined by end points. It is astonishing that even the addition of a semicolon to segmental boundaries is sufficient to improve machine translation. This also explains the central problem: machine translation does not fail because of mixing up questions and statements, but because of mixing up segmental units and enjambements.

In our future work, we plan to train a specific system on translating unpunctuated poetry in order to compare the results with manual translations. The fact that we add punctuation signs on the basis of oral representations of the poems is acceptable when it comes to audio poems, in which the oral representation is an essential part of the poem as a piece of art, closely connected to the written form.

Bibliography

[1] Baker, M. (1994): In Other Words: A Course Book on Translation. London, New York: Routledge.

[2] Halliday, M. A. K. (1985): An Introduction to Functional Grammar. London: Edward Arnold.

[3] Han, S.: Free Google Translate API for Python. Available on https://pypi.org/project/googletrans/. Last accessed at 15. August 2019.

[4] Hosseini-Maasoum, S. M. / Mahdiyan, M. (2012): Punctuation in Translation: The Unseen Side of the Coin. Mediterranean journal of social sciences, 3(11):25–32.

[5] Kirkman, J. (2006): Punctuation Matters: Advice on Punctuation for Scientific and Technical Writing. Routledge study guides. Routledge.

[6] Kobus, N. (2006): Hard cover: Gedichte. Ardey Verlag, Münster.

[7] Lyrikline Literaturwerkstatt Berlin: Lyrikline: listen to the poet. Available on www.lyrikline.org. Last accessed at 03. September 2019.

[8] Newmark, P. (1988): A Textbook of Translation. Prentice Hall.

[9] Papineni, K. / Roukos, S. / Ward, T. / Zhu, W-J (2002): BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL ’02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

[10] Shiyab, S. M. (2017): Translation: Concepts and Critical Issues. Garant Publishers.

[11] Wesling, D. (1996): The Scissors of Meter: Grammetrics and Reading. University of Michigan Press.