A translation-oriented categorisation of wordplays

Michel Delarche[a]

[a] *Université Paris-Diderot UFR EILA 8, place Paul Ricoeur, 75013 Paris, France*

Abstract

We show here that it is possible to pragmatically base a wordplay categorisation scheme on the different tools and algorithms needed to support the (more or less) automated detection and translation of wordplays, especially those whose translation by existing generic tools is impossible or inadequate. This approach also provides a way to develop metrics for quantifying the quality of wordplay translations.

Keywords [1]

wordplay categorisation, automated wordplay detection, wordplay translation

1. Introduction

There exist various categorisation schemes for wordplays based either on a non-academic cataloguing of wordplay types (as enumerated in [1] for example) or on linguistic structures (eg. [2], [3]). In the context of the JOKER project [4] this working note proposes a different approach, whose scope is however limited to alphabetic languages.

Some authors restrict the notion of wordplay to phonetic, semantic and/or syntactic wordplays based on polysemy or homophony (informally known as puns). In this note, we address a wider variety of wordplays. We also suggest additional tools or algorithmic filters for specific subtypes of wordplays.

The deterministic algorithms described in this note should be viewed as both a description of the heuristic processes at work in human translation activities and an outline specification of heuristic tools aimed at facilitating the detection and translation of wordplays too complex to be confidently delegated to automated translation systems functioning as blackboxes. In this respect, we agree with Miller [5] in respect of the many pitfalls plaguing a fully automated approach to wordplay translation. However, by contrast with his PunCAT tool (an ongoing development presented in [6]), our approach, if implemented, would rely on simple searching and matching algorithms operating on (supposedly digitised) traditional dictionaries rather than on database models and engines.

This note is divided into sections and sub-sections reflecting different structural types of wordplays. Each (sub-)section contains:

1. a definition of the type of wordplay it addresses,
2. a list of tools needed for automating the detection and translation of the wordplays,
3. an informal description of the algorithms to be used,
4. a short discussion of quality assessment metrics.

2. Some letter-based constructs

Certain wordplays are based on selections, permutations, repetitions or suppressions of letters: acronyms, acrostics, lipograms, palindromes, pangrams...

In this section we compare two structurally similar wordplays, since they are based on the selection of first letters from words: acronyms and acrostics.

1CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: delarche@noos.fr
ORCID: 0000-0003-3710-8165 (A. 1);

2.1. Acronyms

Acronyms are utterable pseudo-words (or real words) consisting of the first letters of a sequence of words. In this note focused at wordplays, we exclude those acronyms that have been turned into common words, like 'radar' (radio detection and ranging) or 'scuba' (self-contained underwater breathing apparatus) or that are never translated like "the SALT agreements" ("les accords SALT") or "the BBC" ("la BBC"). Acronyms for international organisations admit internationally standardised translations like NATO=OTAN or WHO=OMS and are not concerned either.

When acronyms are meant as wordplays, like the TWIT ("True Worshippers of the Ineffable Tetractys", a pseudo-Pythagorician sect invented by Thomas Pynchon in [7]) or the MOU ("Mouvement Ondulatoire Unifié" a mock political party created by the French humourist Pierre Dac [8] in 1965) they are always written in capital letters (sometimes separated by full stops), which makes their detection extremely simple, providing the established acronyms described in the previous paragraph be filtered out. There may be some semantic relation between the stand-alone meaning of the acronym and its developed form, but is it not necessarily the case (as shown by the ALICE acronym below, certain acronyms may sound playful *per se*. independently from the meaning of their developed form.)

The translation of acronyms requires the following tools: a a source-to-target language dictionary (STLD) to determine the meaning of the acronym and also of its component words in the target language. A dictionary of names (DoN) would also be necessary because some acronyms are based on names, like the mnemonic ALICE (Alert, Lockdown, Inform, Counter, Evacuate).

To initiate the translation process, a list of acceptable translations in the target language for each component word can be determined e.g. in the case of TWIT, with French as the target language, Vrai(e), Véritable, Authentique, Certain(e), Incontestable, Juste, Loyal(e), Réel(le) etc for 'True', Servants, Adorateurs, Disciples, Sectateurs, Dévots, Croyants, Zélateurs etc for 'Worshippers' etc.). 'Ineffable' does not need to be translated, neither does 'Tetractys'. Then, starting from a meaning of the acronym (IDIOT or ABRUTI for TWIT and SOFT or LIMP for MOU) a candidate selection algorithm would try to match one or more acceptable translations with the successive letters of the target algorithm: Incontestables Dévots/Disciples de l'Ineffable Tetractys provide 4 matches out of the 5 needed for IDIOT, which is a good starting point for the human translator (selecting some adjectives starting with O, like Optimal or Original, would provide some additional assistance). For ABRUTI, our lists of synonyms would allow for a partial correspondence only: Authentiques B R U du Tetractys Ineffable.

Translating MOU into SOFT with the same synonym-matching algorithm would pick matches from three lists, the first one containing Association, League, Move, Organisation, Party etc. for 'Mouvement', the second one Fluctuating, Ondulating, Oscillating, Tremulous, Wavy etc. for 'Ondulatoire' and the third one Coalesced, Federated, Integrated, Merged, Unified, United etc. for 'Unifié'. Matching OFT with Organisation of Federated Tremulations and LIM with League of Integrated Moving would each provide a useful starting point: the end points could be Social/Special Organisation of Federated Tremulations and League of Integrated Moving Parties, respectively. (The resulting acronyms are baroque assemblies of words, very much like the original ones: achieving fully rational meaningfulness is not really essential in this kind of exercise.)

As implicitly shown by the above proposed algorithm, the quality criteria of the translation of an acronym are:

- respecting the semantics of the source acronym,
- minimising the difference in length between the source and target acronyms,
- respecting the semantics of the developed acronym (allowing for some permutations and/or derivations of the target words).

It is now possible to build up quantitative metrics reflecting these three criteria. For example, allocating N points for fully respecting the semantics of a source acronym of length N, adding 1 point for each target word matching a source word and subtracting 1 point for each non-matching target word would linearly aggregate the three criteria, with a higher weight granted to the meaning of the whole acronyms. Applying this metric to IDIOT vs TWIT, we get: 4 source letters + 4 words matching the source words - 1 non matching word = 7.

For ABRUTI we get only: 4 source letters + 3 matching words - 3 non-matching words = 3. According to this metric, IDIOT is a better translation of TWIT than ABRUTI.

2.2. Acrostics

An acrostic is a poem (resp. a piece of prose) where the selection of the first letters of successive verses (resp. sentences or paragraphs) produce either a single word (frequently the name of the addressee, like ELIZABETH in an acrostic by Edgar Allan Poe [9], but it may be a common noun or infinitive verb form) or a multi-word expression. We denote these two subtypes of acrostics as name acrostics and multi-word acrostics, respectively.

2.2.1. Name acrostics

The only tool required for the detection of name acrostics is a dictionary of names (DoN). Building up the relevant letters into a string of characters and then checking the result is so simple it can hardly be termed an algorithm.

Translating name acrostics would require an additional tool: a STLD. The translation algorithm would start from the word in the source text associated with the first letter of the name, then find its translation(s) into the STLD. If a translation starts with the same letter, we have a good candidate. We can look for the translations of subsequent words in the first verse to find other candidates. For example, in a translation from English into French, if the name is IRIS, 'Intense' as the first word of the first verse would be an obvious candidate since it also exists in French. If we find the word 'here' in the rest of the verse, 'Ici' would be another candidate for the French version.

Repeating the process for each verse, we get a set of possible starting points that would be heuristically helpful to the human translator. As regards the quality metrics for a name acrostics, respecting the name is the key quality factor (replacing IRIS by IRMA would get only a 50% mark).

This simple metric could be refined by comparing the selected target candidates with the source text: for example, 'Ici' would be a better candidate than 'Incongru' for translating an acrostic starting with 'Iffy'.

2.2.2. Multi-word acrostics

Multi-word acrostics would require a more combinatorial algorithm to determine the sequence of words by splitting the acrostic string at all the right places (starting with a list of potential first words, then a list of potential second words for each potential first word and so on). At the end of the process a shortlist of solutions can be proposed to the user. Including some syntactic rules, e.g. that in Western European languages "*definite determiners (the, le, la, les, der, die, das, el, los, gli, etc.) should be followed by a noun or an adjective*" in the algorithm would further narrow the list of plausible results.

Translating a multi-word acrostic would involve the same kind of candidate selection algorithm as previously described for name acrostics. The quality metrics that may be developed for multi-word acrostics are also the same.

We can see that, by comparison with name acrostics, the detection of multi-word acrostics requires an additional tool (an SLMD). It also requires a specific string segmentation algorithm.

In our approach, these two differences justify the use of distinct categories for these two types of acrostic. However, the candidate selection algorithm and the STLD to be used for supporting the translation process are the same for both.

3. Some polysemic constructs

Polysemic wordplays rely on some polysemic keyword(s), that is a word or a set expression that can have two (or more) different senses. In such wordplays, the keyword or set expression is used in a context where its different possible senses constitute the wordplay. The script-based General Theory of

Verbal Humour developed by Attardo and Raskin [10] proposes a formal model for characterising such semantic disjunctions and Low [11] has listed the generic strategies used by translators (including the ultimate option of leaving a pun untranslated); he also described step-by-step transformation mechanisms operating in the source language and/or the target language so as to support a systematic approach to the translation process by finding semantic approximates.

3.1. Single pivotal keyword

The structure of these wordplays involves a single pivotal keyword, for which the upstream context points at one of the possible meanings while the downstream context leads to another interpretation:> "I took several sick leaves last year, because the trees were suffering from the drought."

The pivotal word is 'leaves'. The beginning of the sentence seems to be about taking a sick leave several times, while the end of the sentence makes sense only if more than one sick leaf has been gathered from the trees. The same structure can be instantiated through a short dialogue featuring a non sequitur:

"- Je suis allé à la boulangerie, et j'ai pris de la brioche.

- Pour t'en débarrasser, fais un peu de sport !"

In a lazy English translation, the answer seems irrelevant:

"- I went to the bakery, and I took some brioche.

- To get rid of it, do a little sport!"

Here 'brioche' is the pivotal word with a dual interpretation. This is because the sentence: "j'ai pris de la brioche" can be interpreted colloquially as: "I've developed /a paunch/a corporation/" "I've got a bit of a tummy" (these three translations are provided in [12]) Since the colloquial expression "prendre de la brioche" can be found in an STLD, the double entendre can be detected by scanning the upstream context, as 'bakery' points at the proper meaning of the word 'brioche'.

After detecting 'brioche' as the pivotal keyword, it is possible to check which translations of this pivot may accept several interpretations; neither 'tummy' nor 'paunch' meet this requirement. but 'corporation' would allow the human translator to devise some approximate equivalence, since the word 'corporation' can designate either a commercial company or a potbelly:

"- I started a small business then I developed a corporation.

- To get rid of it, do a little sport!"

This kind of creative translation cannot be obtained from the deterministic algorithms we are discussing here, but identifying automatically the word 'corporation' as a candidate pivotal word in English would be quite helpful to the human translator. The quality criteria (and hence the quality metric) for the automated part of the process would be:

- the correct identification of the pivotal keyword and
- the identification of good candidates for the role of pivotal keyword in the translation (at least one of the two meanings involved in the wordplay should be kept by each candidate).

3.2. Repeated keyword with different meanings

Another construct consists in repeating a word in a sentence with a different meaning. Assigning a correct meaning to each occurrence is also a matter of context analysis.

The identification of repeated words is easy. To decide whether a repeated word can be listed as a potential polysemic keyword, the first tool to be used is an SLMD. If the dictionary contains different possible meanings for the word, the next step consists in determining whether the two occurrences are associated with a single meaning or not.

In some cases, the context allows for the immediate determination of adequate translations:

"L'eau potable c'est bien, mais un vin potable, c'est mieux."

Here, we have two occurrences of 'potable' which can be considered as the repeated keyword for this semantic wordplay. In this sentence, the first occurrence of 'potable' is part of the set expression "eau potable" (drinking water), while the second one has the figurate meaning of 'acceptable', 'passable',

'palatable'. The translation of "eau potable" as "drinking water" may be directly available in the STLD (as is the case in [12]), and the wording "drinkable water" may be concurrently available too.

Selecting the best translation for "vin potable" consists in reflecting as best as possible the dual meaning of 'potable', and 'drinkable' would be the obvious choice. Using 'drinkable water' would not be erroneous, but associating "drinking water" to "drinkable wine" would capture more accurately the humour in the French wordplay. In that case, the selection algorithm could provide both options for the human translator to make his choice.

"When the lockdown parties scandal erupted, the conservative party was nicknamed the party of parties".

This sentence features 4 occurrences of the word 'party', which can therefore be identified as the keyword in the semantic wordplay due to its polysemy.

Here we have two pairs of occurrences of the noun 'party' the singular ones referring to a political party, and the plural ones denoting festive gatherings. Here, a basic automatic translation process may yield erroneous results *inter alia* because, in many other contexts, "the party of parties" means a superlative feast (as it is a construct similar to such expressions as: "the King of Kings" or "the problem of problems").

However, associating 'party' to the preceding occurrence 'conservative party' and 'parties' to 'lockdown parties' would select the correct interpretations. And if the context proposed instead some synonyms of these contextual occurrences such as 'the bring-your-own-booze gatherings' or 'the Tories', using a SLMD for exploring the definition part of the corresponding entries, would allow to associate 'party' and 'parties' to their respective meanings.

In French, 'partie' means 'part', but is also used either in the domain of games, sports and leisure ("une partie de ping-pong", "une partie de chasse") or to denote the genitals ("private parts") in popular language ("il a reçu le ballon dans les parties"). There are also some set expressions like "surprise-partie", "partie fine" whose meanings correspond to the festive gatherings evoked in our example.

"Le parti conservateur a été surnommé le parti des parties fines" would be a good enough translation. An expression like "partie fine" is unlikely to be given in a bilingual translation dictionary but may be found in a dictionary of idioms (DoI).

The detection and translation of such wordplays requires both a SLD and an STLD. An extensive Target Language Dictionary (TLD) for finding synonyms and a DoI for detecting colloquial set expressions in either language may also be necessary.

4. A few preliminary conclusions

We have shown that name acrostics and acronyms would fall in the same algorithm-based category, while multiword acrostics would have to be categorised separately (because their require a specific string-to-word segmentation algorithm and an SLMD).

We have also reached the conclusion that the structural difference between a polysemic wordplay based on a single pivotal keyword and a polysemic wordplay based on the repetition of a keyword need not be ascribed to distinct algorithm-based categories.

These examples demonstrate that non-trivial wordplay categories can be based on different sets of dictionaries and deterministic detection and translation algorithms.

It must also be noted that traditional mono- or bilingual dictionaries include most of the linguistic functionalities we have here identified separately in relation with the type of algorithmic processing envisaged: a traditional paper dictionary is simultaneously a semantic domain dictionary (plus some register indications by means of asterisks or other notations), a pronouncing dictionary and also in part a dictionary of synonyms and a dictionary of idioms; it frequently contains annexes describing conjugations or declensions. An approach of computer-aided translation based on the use of matching algorithms operating on digitized versions of those traditional dictionaries might be of interest, especially for the little used languages that are unlikely to ever benefit from the construction of advanced linguistic databases.

For the sake of brevity, we have not addressed phonemic puns in this paper. In our approach, they would constitute another group of wordplay categories because they would require additional tools such

as pronouncing dictionaries. For puns based on phonemic approximates rather that pronunciations that are strictly the same, more flexible matching algorithms would be needed too.

5. Acknowledgements

Many thanks to Liana Ermakova for inviting me to submit this note.

6. References

[1] Willard R. Espy The Game of Words (1972) Random House ISBN 0-448-01196-4

[2] Giorgadze, M. Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia Linguistic features of pun, its typology and classification ; European Scientific Journal November 2014 /SPECIAL/ edition vol.2 ISSN: 1857 – 7881 e-ISSN 1857- 7431271

[3 Ermakova, Liana, Tristan Miller, Fabio Regattin, Anne-Gwenn Bosser, Élise Mathurin, Gaelle Le Corre, Sílvia Araújo, et al. "Overview of JOKER@CLEF 2022: Automatic Wordplay and Humour Translation Workshop." In Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), edited by Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro, 25, 2022.

[4] Ermakova, L., Miller, T., Puchalski, O., Regattin, F., Mathurin, É., Araújo, S., Bosser, A.-G., Borg, C., Bokiniec, M., Corre, G. L., Jeanjean, B., Hannachi, R., Mallia, Ġ., Matas, G., & Saki, M. (2022). CLEF Workshop JOKER: Automatic Wordplay and Humour Translation. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, & V. Setty (Eds.), Advances in Information Retrieval (Vol. 13186, pp. 355–363). Springer International Publishing. URL: https://doi.org/10.1007/978-3-030-99739-7_45

[5] Tristan Miller "The punster's amanuensis: The proper place of humans and machines in the translation of wordplay" in Proceedings of the Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019), pages 57–64, September 2019. DOI: 10.26615/ issn.2683-0078.2019_007.

[6] Waltraud Kolb and Tristan Miller. "Human–computer interaction in pun translation." In James Hadley, Kristiina Taivalkoski-Shilov, Carlos S. C. Teixeira, and Antonio Toral, editors, Using Technologies for Creative-Text Translation. Routledge, 2022. To appear.

]7] Th. Pynchon, Against the Day (2006) - Penguin Press HC ISBN 9781594201202

[8] P. Dac, Le parti d'en rire : Pierre Dac président ! (2017) - Cherche Midi Ed. ISBN 9782749145266

[9] E. A. Poe, Elizabeth (1829 ca) in The Complete Poetry of Edgar Allan Poe (1996) - Signet Classic ISBN 9780451526403

[10] Attardo, Salvatore and Raskin, Victor. Script theory revis(it)ed: joke similarity and joke representation model" , vol. 4, no. 3-4, 1991, pp. 293-348. URL: https://doi.org/10.1515/humr.1991.4.3-4.293

[11] Peter Alan Low. "Translating Jokes and Puns". In: Perspectives: Studies in Translation Theory and Practice 19.1 (2011), pp. 59–70. ISSN: 0907-676X. DOI: 10.1080/0907676X.2010.493219.

[12] B. T. Atkins, A. Duval, R. C. Milne & al. Dictionnaire Français-Anglais et Anglais-Français (1978) SNL Le Robert, Paris ISBN 2850360082