Published in: Mitkov, Ruslan (Ed.): EUROPHRAS 2017. Computational and corpus-based phraseology: Recent advances and interdisciplinary approaches. Proceedings of the Conference Volume II (short papers, posters and student workshop papers) November 13-14, 2017 London, UK. - Geneva: Editions Tradulex, 2017. Pp. 45-52

Corpus Linguistic Exploration of Modern Proverb Use and Proverb Patterns

Kathrin Steyer

Institut für Deutsche Sprache, R5, 6-13, 68161 Mannheim, Germany
steyer@ids-mannheim.de

Abstract. In my talk, I present an empirical approach to detecting and describing proverbs as frozen sentences with specific functions in current language use. We have developed this approach in the EU project 'SprichWort' (based on the German Reference Corpus). The first chapter illustrates selected aspects of our complex, iterative procedure to validate proverb candidates. Based on our corpus-driven *lexpan* methodology of slot analysis I then discuss semantic restrictions of proverb patterns. Furthermore, I show different degrees of proverb quality ranging from genuine proverbs to non-proverb realizations of the same abstract pattern. On the one hand, the corpus validation reveals that proverbs are definitely perceived and used as relatively fixed entities and often as sentences. On the other hand, proverbs are not only interpreted as an interesting unique phenomenon but also as part of the whole lexicon, embedded in networks of different lexical items.

Keywords: Proverb; Patterns; Corpus Driven Approach

1 Introduction

The proverb lives on. A quick glance at corpora or the world of new media is enough to see the endless creativity of fixed sentences. New media revive one of the oldest (not old-fashioned) text types. These colloquial forms of communication resemble spoken language, and perhaps this is why they seem to inspire the speakers/writers to use them to reflect upon their messages and to comment on problems of everyday life, politics, culture and sports.

As a phenomenon of the mental lexicon, proverbs are also of interest to a branch of linguistics that focuses on language structures and functions, as a phenomenon of the mental lexicon, embedded in networks of constructions and relationships with other lexical items.

EUROPHRAS 2017, pages 45–52,
London, UK, November 13-14, 2017.
https://doi.org/10.26615/978-2-9701095-2-5_006

Language technology, large data collections and sophisticated automatic methods, allow the exploration of current proverb use based on authentic language mass data in a new dimension.

In this paper, I discuss an empirical approach to describing proverbs as frozen sentences with specific functions in current language use, namely corpus-based (see 2.) as well corpus-driven (see 3.) [1] [2]. Corpus analysis helps to conceive their meaning and usage more accurately and nuanced than by pure introspection. Below, I will focus on proverb status and patterns. First, I present our complex, iterative, corpus-based procedure to validate proverb candidates and in the second part, I show how the *lexpan* methodology enables scholars to detect functional restrictions of sentence patterns and the nature of slot fillers.

2 Iterative empirical approach for proverb validation

2.1 The project background

Our proverbial studies have arisen from the multilingual EU project 'SprichWort. Eine Internetplattform für das Sprachenlernen' (Proverb. An online platform for language learning)[3][1]. The goal was to examine and describe the similarities and differences of contemporary proverb use in different languages and cultures, based on exhaustive corpus studies.

The results are documented on the online platform *SprichWort* (since 2012 hosted at the Institute for the German Language (IDS) in Mannheim, Germany).[2] The platform contains of three sections:

a) A multilingual lexicographic database of 300 proverbs in five languages
b) A series of didactical exercises for teachers and students
c) A proverb community – proverbs in social networks like Twitter and Facebook.

For the source language German, we had to run the empirical validation for a total of 2000 German proverbs based on the German Reference Corpus (DeReKo)[3]. The proverb candidates were extracted from dictionaries, textbooks and collections. As a result, about 900 proverbs could be proved in the corpus (this is 45%).

---

[1] This project was financed by the EU commission for two years. Aside our own IDS group, the partners came from Austria, Czech Republic, Hungary, Slovenia and Slovakia (2008-2010, 143376-LLP-1-2008-1-SI-KA2-KA2MP) [4] [5].

[2] The German Database is integrated in the OWID proverb online dictionary that we compile continuously [6].

[3] The German Reference Corpus, located at the IDS is the world's largest linguistically motivated collection (31,68 billion words: March 2017) of electronic corpora with written German texts from today and the recent past [7]. For this paper, I used a sub corpus, the W archive (W) with a size of about 9 billion words. All queries (Q) are formulated in COSMAS II, the Corpus Search, Management and Analysis System in DeReKo. The frequencies are in brackets.

In the absence of a comprehensive corpus validation in the past, we first had to develop an iterative methodology. Searching for proverbs in a general language corpus is not a trivial task that could be allocated solely to a machine. Few if any assumptions about the surface form and behavior of a proverb can be made in advance (a priori), because again and again corpus evidence proved our intuition wrong.

Considering established proverb definitions [8] [9] we had to address the following questions:

1) Is a candidate indeed a lexicalized sentence or a non-finite clause or merely a phrase?
2) Are well-known quotations or new sentences, e.g. advertising slogans, also used frequently as wisdom sentence without the original context of creating?

2.2 Corpus-based validation of the proverb status

2.2.1 Frozen sentence or phrase?

As I mentioned in 2.1., the corpus-based validation is a complex quantitative-qualitative procedure. Each proverb candidate must be examined individually in an alternation of automatic analysis and formulation of hypotheses. If one searches for a fixed sentence structure in the corpus, one will only find this sentence. All possible variations, extensions and reductions will not be covered by this search. Therefore, the best strategy is to start with a wide search which is then gradually restricted.

The first step was to check whether the lexical components of the proverb candidate appear in the same sentence at all. If the search for the lexical components in the same sentence was successful, this can indicate a proverb. KWIC concordance lines help to quickly check the important proverb criterion: Is the form a sentence or a non-finite clause equivalent to a sentence [10]? In some cases, the pure search of two components in a sentence is sufficient to get a positive result like in the following example of the proverb candidate *Not macht erfinderisch*[4] (Distress makes ious[5]).

(1) P14 *Not macht erfinderisch*. Die Reform beginnt im Kopf.
RHZ01 *Not macht erfinderisch*: Die Soldaten hatten Mangel an Schreibpapier.
NUN06 Allein in Bayern fehlen zurzeit 30 000 Ausbildungsplätze. Die *Not macht erfinderisch*, auch auf Arbeitgeberseite.

In other cases, one has to reduce the query to a very narrow range, e.g. for the proverb candidate *Zeit ist Geld* (Time is money)[6]: In this case, wide queries are not useful because of 'false hits', e.g. *Die Fahrer hören nichts vom Lärm* (The drivers hear nothing of the noise) or *Zeit und Geld* (Time and money).

---

4 Q: „Not" /s0 &erfinderisch (2.973)
5 The literal translations are bracketed; semantic equivalents are marked by ee.
6 Q: Zeit /+w1:1 ist /+w1:1 Geld (1.976)

There also were surprising observations, particularly in relation to those candidates that seemed to be common proverbs based on our intuition. The analysis, however, did not support this. One example is the candidate *Niemand ist ohne Fehl und Tadel* (Nobody is without faults and blame, ee: Nobody is perfect). *Fehl* is a Middle High German word (meaning: mistake, weakness).

(2) F95 Wer *ohne Fehl und Tadel* ist, der werfe den ersten Stein.
N92 löste diese Aufgabe *ohne Fehl und Tadel*.
M98 selbst Heilige sind nicht frei von *Fehl und Tadel*.

(He who is without *faults and blame*, may throw the first stone
… did the task without *faults and blame*
Even saints are not free from *faults and blame*)

In this case, only the word pair *Fehl und Tadel*, mostly in combination with the preposition *ohne* (without) is fixed, but the contexts vary. It is a very frequent multi-word expression (2.235), but not a proverb.

2.2.3 Quotation or proverb?

One of the common sources for the genesis of proverbs is quotations or citations. To be regarded as a real proverb, a sentence has to be used frequently in daily communication in several situations and also without reference to the original quotation context. Let us take a look of the proverb candidate *Viel Lärm um nichts* (Much noise about nothing, ee: Much ado about nothing). Many occurrences related to the comedy of William Shakespeare. We tried to find as many words as possible that indicated a Shakespearean context in any form. The final, very complex search query excluded such thematic words or tokens like *Shakespeare*, *Hollywood*, *Branagh* or Germans words for 'comedy', 'clock', 'movie', 'stage direction' and 'actor' or compound words with *theatre* etcetera. [7] This search reduced the frequency from 4.455 to 2.681 really 'Shakespeare free' occurrences and reflect its usage as a proverb.

This double life as a quotation and a proverb is a very frequent phenomenon in the corpus.

Currently, we also use corpus-driven methods to discover new proverb candidates, among others collocation profiles of proverb introducers or labels like *proverb*, *wisdom* or *as the saying goes* and proverbial keywords, e.g. *world, time* or *you shouldn't*.

At the end of this chapter I will make a brief note about *proverb frequency*: Calculating proverb frequency is a complex problem which has no standard solution. There

[7] Q: (&viel /s0 Lärm /s0 "nichts") %s1 (&Shakespeare oder &Komödie oder &Hollywood oder &Uhr oder &Stück oder &Schauspieler oder &Theater oder &Kino oder &Film oder &Verfilmung oder &Regisseur oder Theater* oder *theater oder Branagh* oder ado oder ZDF oder ORF oder 20.00 oder 20.15 oder *komödie oder Sommernachtstraum oder Sa oder CET oder Zitadelle) (2.681)

will be different results depending on the corpus and the search query that was used. Statements about frequency are only meaningful if one clearly outlines on which corpus basis and with which search queries the numbers have been obtained. It is also recommended that reference is made to proportional frequencies or frequency trends rather than absolute numbers.

3 Proverb patterns – corpus driven

3.1 Slot filler analysis with *lexpan*

The corpus-driven exploration of proverb schemes, so called 'proverb patterns' [11][8], is one of the most innovative areas in paremiology. The results can also contribute to new researches in pattern-based Phraseology, Construction Grammar and Cognitive Linguistics. Proverb patterns consist of fixed lexical components ('lexical anchors') as well as slots. These fillers indicate realizations of those patterns in specific communicative situations, both proverbial and non-proverbial. They can be determined by different degrees of typicality (frequencies). Fillers have similar semantic and/or pragmatic characteristics, but don't necessarily belong to the same morpho-syntactic category. The nature of filler groups cannot be predicted a priori or by rules but only based on an inductive, bottom-up analysis. For this, we developed the language independent pattern matching tool *lexpan* [13] (free available since March this year). *lexpan* makes it possible to explore corpus data in its own working environment independent of a corpus platform. It can be used for restructuring and annotating the interpreted data and for visualizing it for new forms of lexicographic representation. Currently, we can work with collocation data and KWIC lines.

The KWIC lines have been captured by a search pattern with fixed lexical elements and slots (one or more), and we can observe the proportional relations of the fillers and of the underlying syntagmatic structures.

3.2 Semantic and functional restrictions of proverb patterns

I will demonstrate the approach using three examples. The first pattern is *Andere* X, *andere* Y (Other X, other Y)[9]. Table 1 shows a *lexpan* filler table of the X *and* Y slots (counted as bigrams):

[8] In Folklore, the idea of proverb patterns and underlying abstract meanings and functions dates back to the 19th century. Overviews are given by Röhrich & Mieder [7] and Mac Coinnigh [12].

[9] Q: $andere /+w2:2 $andere (5.715)

filler	frequency	%	comment
Länder ... Sitten	1089	21,24	(countries – customs)
Zeiten ... Sitten	110	2,15	(times – customs)
Stimmen ... Räume	54	1,05	(voices – places)
Räume ... Träume	22	0,43	(places – dreams)
Völker ... Sitten	20	0,39	(peoples – customs)
[…]			
Länder ... Kulturen	13	0,25	(countries – cultures)
Länder ... Regeln	12	0,23	(countries – rules)
Sprache ... Kultur	12	0,23	(language – culture)

Tab. 1. Automatic filler tables of *Andere* X, *andere* Y (lexpan snippet)

In 21.24 % of all occurrences the fillers are *Länder* (countries) and *Sitten* (customs) for the common German proverb *Andere Länder, andere Sitten* (Other countries, other customs). This result is typical for many patterns: Often proverbs are the prototypical realizations. Because of that, one can assume a single proverb entry in the mental lexicon. An interesting observation is that the majority of fillers in the X position refer to concepts of nationality in a broad sense, the fillers in the Y positions refer to norms of behavior (also in the range of low frequency). In about 30% the X filler is *Länder*. Therefore, this pattern is highly restricted by concepts like nationality and behavior.

My next example treats the pattern *Niemand ist* X (Nobody is X). The X slot is filled by a number of adjectives, but there are only two semantic groups that indicate a lexical pattern quality: PERFECTION and REPLACEABILITY. Table 2 illustrate these filler groups, qualitatively systematized by the *lexpan* feature for manual annotation:

filler	frequency	%	tag	comment
perfekt	171	15,23	[[PERFECTION]]	perfect
unfehlbar	52	4,63	[[PERFECTION]]	infallible
unschlagbar	27	2,40	[[PERFECTION]]	unbeatable
vollkommen	20	1,78	[[PERFECTION]]	perfect
fehlerfrei	10	0,89	[[PERFECTION]]	faultless
immun	10	0,89	[[PERFECTION]]	immune
unersetzlich	56	4,99	[[REPLACEABILITY]]	irreplaceable
unersetzbar	21	1,87	[[REPLACEABILITY]]	irreplaceable
unantastbar	8	0,71	[[REPLACEABILITY]]	untouchable
sakrosankt	5	0,45	[[REPLACEABILITY]]	sacrosanct

Tab. 2. Grouped filler tables of the pattern *Niemand is* X (Nobody is X)

The prototypical realization of this pattern is again a proverb: *Niemand ist perfekt* (Nobody is perfect). All these fillers indicate expressions of worldly wisdom. These

distinct characteristics become clear when searching the past tense of the sentence: *Niemand war* X (Nobody was X). In this form, no German adjective refers to one of the two concepts (**Niemand war perfekt / unersetzlich*). The other realizations are not expressions of wisdom but regular declarative sentences, e.g. *Nobody was injured / responsible / sad*. Of course, these sentences can also be contextualized pragmatically but this is not incorporated in the meaning of adjectives like *injured* or *responsible*.

A further insight of our exploration is that there exist transition zones range from genuine proverbs to bogus proverbs to non-proverb realisations of the same pattern, e.g. ***Wer*** X ***der*** Y (***He who*** X Y). The most frequent and prototypical realizations are also the proverbs ***Wer*** *rastet,* ***der*** *rostet* (**He who** rests, grows rusty), ***Wer*** *sucht,* ***der*** *findet* (**He who** searches, finds) or ***Wer*** *wagt,* (***der***) *gewinnt* (**He who** takes risks, wins). The second group are realizations that seem like proverbs because of their typical short proverb structure, but the filler of the X Y slots are strongly context-dependent and often rare: ***Wer*** *fastet,* ***der*** *friert* (**He who** fasts, freezes) or (***Wer*** *kämpft,* ***der*** *tötet*) (**He who** fights, kills). One can also find non-proverb realizations which are regular uses of the pattern without the fixed proverb structure: ***Wer*** *60 von 74 Punkten erreicht,* ***der*** *hat bestanden* (**He who** reaches 60 of 74 points, has passed), ***Wer*** *es etwas gemütlicher mag,* ***der*** *ist beim Freizeitrudern richtig* (**He who** likes it more relaxed, is at the right place with recreational rowing). Independent of the structure all these realizations still transport the same holistic meaning: 'If certain facts are true for a person, in consequence the other fact must also be true for him/her'.

In future, proverbs patterns will feature as a new user approach in our dictionary.

4 Conclusion

Our corpus linguistic exploration of modern proverb use shows on the one hand that proverbs themselves can be realizations of more general patterns and schemes, furthermore, they share attributes and characteristics with non-proverb multi-word units and other lexical items. It is assumed that there are two lexicon entries: once as a lexicalized proverb and once as a pattern that can also be activated for non-proverb use.

On the other hand, our exploration proved that proverbs are definitely perceived and used as relatively fixed entities and often as sentences. Speakers seem to have a strong sentence-level knowledge, even though they do not distinguish proverbs from sayings, mottos etc. This sentence-level knowledge enables them to create analogies and to produce new realizations of the same proverb pattern. Proverbs are more salient in the mind of the speakers, while non-proverb units of the same schema tend to be subject to creative ad-hoc variations.

This raises the interesting question for future research why some proverbs have hardly any variants while others have many. As you can see, strictly corpus-based proverb studies can create a fresh impetus for a pattern-based theory of the lexicon and vice versa.

References

1. Sinclair, J.: Corpus, Concordance, Collocation. University Press, Oxford (1999).
2. Hanks, P.: Lexical Analysis. Norms and Exploitations. MIT Press, Cambridge (MA) (2013).
3. Sprichwortplattform, http://www.sprichwort-plattform.org/, last accessed 2017/08/27.
4. Steyer, K. (ed.): Sprichwörter multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömiologie. Narr, Tübingen (2012).
5. Ďurčo P., Steyer, K., Hein, K.: Sprichwörter im Gebrauch. Unveränderter Wiederabdruck der 2015 in Trnava erschienenen Erstausgabe. Institut für Deutsche Sprache, Mannheim (2017).
6. OWID Sprichwörterbuch, http://www.owid.de/wb/sprw/start.html, last accessed 2017/08/27.
7. Institut für Deutsche Sprache: *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-I* (Release vom 08.03.2017). Mannheim, www.ids-mannheim.de/DeReKo, last accessed 2017/08/27.
8. Röhrich, L., Mieder, W.: Sprichwort. Metzler, Stuttgart (1977).
9. Hrisztova-Gotthardt, H., Aleksa Varga, M.: Introduction to Paremiology: A Comprehensive Guide to Proverb Studies. Berlin, de Gruyter (2015).
10. Lüger, H.-H.: Satzwertige Phraseologismen. Eine pragmalinguistische Untersuchung. Praesens, Wien (1999).
11. Steyer, K.: Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht. Narr, Tübingen. (2013).
12. Mac Coinnigh, M.: Structural Aspects of Proverbs. In: Hrisztova-Gotthardt, H., Aleksa Varga, M. (eds.), pp. 112-132.
13. lexpan: Lexical Patterns Analyzer. Ein Analysewerkzeug zur Untersuchung syntagmatischer Strukturen auf der Basis von Korpusdaten – An tool for the exploration of syntagmatic structures based in corpus data http://www1.ids-mannheim.de/lexik/uwv/lexpan.html, last accessed 2017/08/27.