Claudine Moulin/Iryna Gurevych/
Natalia Filatkina/Richard Eckart de Castilho

Analyzing Formulaic Patterns in Historical Corpora

Abstract

This paper aims to point out a linguistic phenomenon that due to the current stage of research can be analysed only insufficiently with the help of an electronic text corpus. In this way, the paper adds a new aspect to the discussion about historical corpora by tackling the question of how they should be designed in order to be useful for linguistic research on so-called *formulaic patterns*. The novelty of the question becomes apparent considering the fact that at present such historical corpora do not exist. In section 1, we define the term *formulaic pattern* because a clear understanding of this phenomenon is a prerequisite condition for collaborative research of it by historians of language and corpus and computer linguists. Section 2 gives a brief outline of the state of the art in the field of modern formulaic language within the framework of corpus and computer linguistics. Section 3 shows that some well known problems in this area are exacerbated when applied to historical texts. Section 4 presents a possible solution that has been implemented by the HiFoS Research Group at the University of Trier (Germany). Joint research efforts planned with UKP Lab at the TU Darmstadt (section 5) demonstrate that the restrictions posed by historical formulaic patterns are challenges to be overcome, rather than insurmountable obstacles.

1. Formulaic patters at the crossroads of (historical) linguistics, corpus and computer linguistics: Looking for a common ground

Formulaic patterns is not yet a well-established linguistic term. The linguistic phenomena it is applied to have been studied mostly within the framework of phraseological research. *Phrasemes, set phrases* or in German *Phraseologismen* are understood here as expressions that comprise a minimum of two words (constituents) and a maximum of a sentence[1]. They are syntactically more or less frozen and can (but need not) be semantically ambiguous or idiomatic (Burger^4 2010: 11-15). For different phrasemes, these semantic and syntactic features are characteristic to a different extent. Therefore, lin-

[1] From a historical point of view, these criteria can only serve as a starting point while identifying formulaic patterns. Cf. Filatkina/Gottwald/Hanauska et al (2009) and section 3 in this paper for more details.

guists distinguish various types of phrasemes: idioms (*to spill the beans, to kick the bucket*), binomials (*in black and white*), collocations (*to give a paper, boiling water*), proverbs (*Clothes make the man*), routine formulae (*Ladies and Gentlemen!, The floor is yours!*), quotations (*To Be or not to Be - That is the question!*) and so on, to name just a few types. Phrasemes are conventionalised expressions reproduced by speakers in their particular structure and meaning. Furthermore, some types of phrasemes are strongly tied to a particular cultural background and transmit cultural elements through their image components (Dobrovol'skij/Piirainen 2005). Since its establishment in the 1940s and its development into an independent, international branch of linguistics in the 1970s, phraseological studies have proven that phrasemes are a universal phenomenon typical in all modern languages but strongly dependant on the communicative and cultural conventions of a given language. Until very recently, phraseological research has mostly addressed modern languages spoken in Europe[2]; its main focus has been semantics and pragmatics.

A significant shift towards the investigation of the structure of phrasemes, their high potential for variation and occasional modifications occurred in the 1990s, partly driven by the advent of corpus and computer linguistics (Fellbaum 2006, Heid 2008). As computer linguistics uses many methods that have evolved from corpus linguistics, we will only use the term computer linguistics for the rest of the paper and implicitly refer to corpus linguistics as well. What is described here with the terms *collocations* (Fellbaum 2007, Evert 2005, Evert 2008) or *multi-word-units* (Tschichold 2000, Sag et al 2001) differs from the above mentioned linguistic definitions. Collocations and multi-word-units are understood as statistically significant co-occurrences of mostly two lexical items. According to this definition, the example *New York* is considered to be a typical collocation in corpus linguistics whereas its consideration as a phraseme is questionable for a phraseologist[3]. Despite these differences, corpus linguistics was one of the first disciplines to prove the high significance of syntactically more or less frozen constructions for human communication.[4] Sinclair's idiom principle (1987), Hoey's lexical priming model (2005) or Wray's concept of formulaic language (2008) build on the basic idea that in a natural communication process

[2] Cf. a different approach in Piirainen (2012).

[3] For more discussion concerning the definition problem cf. recently Sirajzade (2012).

[4] Jackendoff (1995: 156) observes for English: “The lexicon of a language available to speakers in everyday situations contains at least as many multi-word expressions as single words”. Cf. also Sag et al (2001), Fellbaum (2007).

the words of a language exist not in isolation but in a syntagmatic interplay with each other. As a result of this interplay, words evoke meanings attested to them by the members of their respective language community. While phraseological studies open their boundaries to the methods of computer linguistics, computer linguists start addressing the complex semantic nature of phrasemes in text corpora (f.e. disambiguation, literal/non literal meaning; Li/Sporleder 2010; Li/Roth/Sporleder 2010) – the question that used to be central for classical phraseological research (as mentioned above).

As phraseological studies, computer linguistic research was mostly restricted to modern languages. However, the extensive research on historical German texts carried out in the HiFoS Research Group at the University of Trier (Germany) offers evidence that enables a sounder evaluation of some of the criteria used in determining formulaic diction (cf. section 3). Therefore, in the HiFoS-project, we speak about historical *formulaic patterns* rather than *historical phraseology* or *multi-word-units*. This term is more general than *phraseology* and enables scholars with different scientific interests to include even those patterns into the analysis that have a very low degree of syntactical stability and semantic idiomaticity. Formulaic patterns are not necessarily restricted by the length of two words on one hand and by sentence boundaries on the other hand. These patterns can rarely occur in historical texts. Furthermore, in the HiFoS-project, an expression is considered to be formulaic if, in addition to its concise syntactic structure, its pragmatic functions are evident and central to a given text (Filatkina 2009a; Filatkina/Gottwald/Hanauska et al 2009).

2. Formulaic patterns in modern languages as "a pain in the neck" for corpus and computer linguistics

Despite the long tradition of computer linguistic research on formulaic patterns in modern languages, they are still considered to be "a pain in the neck" (Sag et al 2001) from the technical point of view.

First of all, the question of what corpus size is sufficient for research on formulaic patterns remains unanswered. Scholars have already pointed out that it should be substantially larger than corpora used for research on other linguistic phenomena. According to Geyken (2004), even a 100 million words corpus might not be sufficient for formulaic patterns due to the discrepancy in the type-token-frequency: Even though different types of formulaic patterns are highly constitutive for many texts (cf. section 1), the token frequency of each type might be very low (cf. also Claridge 2008). Furthermore, the common practice in corpus linguistics to compile corpora from sporadically chosen longer text excerpts does not appear to be helpful for research on formulaic patterns as the occurrence of formulaic patterns within a text excerpt can never be predicted in advance. One possible solution is to restrict the choice of texts and focus on certain genres. In reality, this approach turns out to be difficult as well because at the current stage of phraseological research little knowledge about the formulaic character of single text genres is available.

Secondly, no annotation standard or formal categorization scheme has so far been developed for formulaic patterns. One reason for this might be the difference in research perspectives between phraseological and computer linguistic research on formulaic patterns: While the former have, until very recently, been mostly semantically oriented, the latter can proceed to semantic analysis only after morphology and syntax have been reliably analyzed.

Thirdly, the (semi)automatic identification and retrieval of formulaic patterns continues to be a challenging task. The existing approaches can be roughly grouped in 1) association measures tools, 2) formal preference tools, and 3) distribution semantic tools. At present, the tools of the first two groups are most common. They are statistical in nature and operate on the basis of statistically significant co-occurrences in shallowly annotated corpora. Association measure tools have proven to be particularly useful in the identification of set phrases with a frozen syntactic structure and of those consisting of not more than two lexical items. Formal preferences include morphological and syntactic restrictions as well as idiosyncratic features in the structure of formulaic patterns but these are less efficient for the identification of syntactically absolutely regular patterns. Absolutely stable patterns, units limited to two lexical items or expressions with morphological and syntactic irregularities are far from being the majority of formulaic patterns. In addition to corpus linguistics, the Natural Language Processing (NLP) has also addressed the question of identifying multi-word expressions. Here, the distinction of literal and non-literal usages was of a particular interest. For distinguishing non-literal idioms, approaches looking at lexical cohesion (global lexical context, local lexical context, discourse context) and to a lesser extend at some syntactic features show good results (Li/Sporleder 2009), particularly as they allow generalization across idioms and address the problem of their low token frequency. At a more abstract level, some similarities can be found between identifying formulaic patterns and the task of relation extraction in Text Mining (Heyer/Quasthoff/Witting 2008). Answering the question how well do different text re-use techniques work in order to link the similar but still variant paraphrased passages (Büchler/Crane/Mueller/Burns/Heyer 2011), related versions of the same text (f.e. different versions of Bible) were required and analysed so far. As variants of formulaic patterns occur not only in related texts, these approaches need to be extended to texts of completely different origins. It is important to note that all tools and techniques mentioned above were developed and implemented on modern and/or normalized texts. Therefore, the development of efficient identification and retrieval tools for historical formulaic patterns remains a future research task so far.

3. Historical formulaic patterns of German and text corpora

A corpus-based study of historical formulaic patterns in German turns out to be an even more challenging task at present. Until very recently, no philologically reliable corpus of historical German texts existed that would allow a systematic and consistent investigation into the dynamics of formulaic patterns from the beginning of the written tradition until the Early New High German period. Recently started projects such as *Referenzkorpus Altdeutsch*[5], *Das annotierte Refenrenzkorpus Mittelhochdeutsch (1050-1350)*[6] at the University of Bonn (Dipper 2010) and *Referenzkorpus Frühneuhochdeutsch* at the Martin-Luther-University Halle-Wittemberg[7] aim to fill this gap but, to the best of our knowledge, they have yet to be investigated from the perspective of formulaic patterns. The major difference between these corpora and the previously existing ones is the fact that they include full texts and not text excerpts, texts of different genres and different authors. These are essential requirements for contemporary corpus-driven and corpus-based studies on historical formulaic patterns as so far they have been limited to just a few sources (cf. the overview in Filatkina 2009b; Filatkina/Gottwald/Hanauska et al 2009). In contrast to previous text collections, these recent corpora are based on diplomatic transcriptions of original texts. This makes them particularly suitable for research into morphological, syn-

[5] http://www.deutschdiachrondigital.de/.

[6] http://www.linguistics.ruhr-uni-bochum.de/~dipper/project_ddd.html.

[7] http://www.germanistik.uni-halle.de/forschung/altgermanistik/referenzkorpus_fruehneuhochdeutsc/.

tactic and lexical variation of formulaic patterns that is hardly possible on the basis of normalized text editions.

The results of the HiFoS Research Group show that variation is the most characteristic feature of formulaic patterns in historical German texts (Filatkina 2009a; Filatkina 2012). In order for language historians to carry out extensive research into variation models and their dynamics, this fact must be taken into consideration while answering the question about the depth of corpus annotation. It also sheds a new light on the process of (semi-)automatic identification of formulaic patterns. These questions have not yet been addressed by any of the existing historical corpus projects. Bennett/Durrell/Scheible/Witt (2010: 65) claim that the identification of formulaic patterns will be the aim at a later stage of the GerManC-project; to our knowledge, no results are available to date.

The decision as to what a formulaic pattern is in a historical text is not trivial even from the point of view of historical linguistics. In a modern language, one can statistically measure the degree of the formulaic diction with the help of sufficiently large text corpora (cf. section 2), questionnaires or interviews. However, these methods of contemporary empirical linguistics are not possible when working historically. Consequently, a researcher is restricted to original texts and singular findings in texts that have happened be handed down from earlier times and whose number is incomplete. One of the more widely accepted criterion for a formulaic pattern is its repetitious occurrence. It would seem a truism that this phenomenon can and indeed must be documented in order to employ the criteria. Thus, for the reasons given above it cannot be put at the centre of linguistic analysis of the historical data. Furthermore, the identification criteria that were established within the framework of phraseological research (polylexicality, syntactic stability, idiomaticity, cf. section 1) often do not apply to historical data where f.e. polylexicality confronts the lack of orthographic norms or the problem of word/sentence boundaries and idiomaticity – the difficulties of hermeneutic interpretation of meaning caused by culture and time distances between present day and historical data. For any given formulaic pattern, it will be necessary to take into account relevant factors which apply to the transmission, geographical space, date of the text and evidence gathered from other languages as well as the cultural-historical role of the expression including not only verbal media but also its visualisations. Corpus and computational studies in the field of formulaic patterns should take these circumstances as a starting point in order to facilitate linguistic research and to take it to a new level. More linguistic knowledge about historical formulaic patterns is required in order to support the corpus compilation and the development of annotation tools and standards, or as Rayson/Piao/Sharoff/Evert/Moirón (2010 44: 2) put it:

> […] it has become increasingly obvious that in order to develop more efficient algorithms, we need deeper understanding of the structural and semantic properties of MWE's [NF: multi-word-expressions], such as morpho-syntactic patterns, semantic compositionality, semantic behaviour in different contexts, cross-lingual transformation of MWE properties etc.

4. HiFoS Research Group

One substantial step towards the collection of such knowledge is the Research Group "Historical formulaic language and traditions of communication" or in German "Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)" at the University of Trier.[8]

What historical research on formulaic patterns has so far been lacking is a systematic investigation of diachrony and synchrony in original historical texts, an investigation into the stability and variation of formulaic patterns in these texts, and the dependency of their usage upon the text genre as well the intertextual specifics of their distribution. These are exactly the goals that HiFoS aims to achieve in order to create a strong disciplinary basis for further research, in particular for collaborative research with other philologies and scholarly disciplines. HiFoS investigates the historical development, stability and variation of different types of formulaic patterns in different German texts over the time period from ca. 700 to ca. 1700 with a strong focus on the oldest – Old High German – texts (ca. 700 to ca. 1050). Several theoretical and methodological principles were established by the HiFoS-Group in order to answer the questions above:

1) In contrast to some other historical projects, HiFoS collects its data from original manuscripts rather than normalized text editions as edited formulaic patterns might have never existed in the edited form in original manuscripts.

[8] The project is financed by the Sofja Kovalevskaja Award 2006 of the Alexander von Humboldt Foundation. For further information cf. http://www.hifos.uni-trier.de as well as the publications of the group members listed on the project website.

2) Due to the current stage of historical corpora compilation and with regard to specific requirements posed by formulaic patterns (see sections 2 and 3) as well as the primary goals of the HiFoS Research Group, great effort was put into manual or rather intellectual extraction, documentation and detailed annotation of such single findings in different texts.

3) As scholarly research generally has little information about what types of formulaic patters can be found in which historical texts and why, the HiFoS project aims to cover German texts of different genres (poetry and fiction, historical legal texts, non fictional sources, f.e. travel reports, religious and pre-scientific texts). The data gained from the fiction texts can be then compared with those formulaic patterns and metalinguistic knowledge about them that are subject of proverbial collections, dictionaries and grammar books[9].

4) Formulaic patterns found manually in original text documents are collected in a database and encoded according to the standards of Unicode and the Text Encoding Initiative (TEI). After the context of the expression in question is noted, its type, different aspects of its morphology, syntax, semantics, pragmatics, and (if possible) cultural historical background, transmission lines and the interdependency of a given formulaic pattern on earlier similar expressions in other languages, particularly Latin and Greek,[10] are analyzed.[11]

5) In order to gain a more or less complete picture of the historical usage and dynamics of a formulaic pattern, the HiFoS database provides the possibility of grouping single findings in a so called *Formulierungstradition*. The collection of variants for one particular formulaic pattern is only one intended way of grouping data. Similar groupings are envisioned for formulaic patterns from specific texts, of specific authors, semantic or pragmatic equivalents and so on.

[9] With this regard, the knowledge gathered within the DoLPh and OldPhras projects becomes particularly beneficial for HiFoS. Cf. http://infolux.uni.lu/phraseologie/ and http://www.oldphras.net.

[10] By doing so, we do not want to draw an immediate conclusion that a given formulaic expression is a loan pattern as this would require an additional systematic research. Such joint research, extensive data exchange and close collaboration were established in 2010 between HiFoS, DoLPh (“Dynamics of Luxembourgish Phraseology”), Aliento (“Analyse Linguistique, Interculturelle d'énoncés sapientiels et Transmission Orient-occident Occident-orient”; http://www.aliento.eu), CASG (“Corpus der arabischen und syrischen Gnomologien”; http://casg.orientphil.uni-halle.de/) und SAWS (“Sharing Ancient Wisdoms”; http://www.kcl.ac.uk/schools/humanities/depts/bmgs/research-section/saw/). At present, the joint research interface and first publications are in preparation.

[11] For the complete description of all areas cf. Filatkina (2009b).

The compilation of extensively annotated data in form of a *Belegkorpus* is one of the major goals of the HiFoS project. But, at the same time, the database is also a research platform that has been already used in several smaller research projects and will be made available online by the end of the project in 2013. The database is crosslinked with an international bibliography about the historical German formulaic language and with the images of original manuscripts available online. This is particularly helpful in situations of a very complex contextualisation of a formulaic pattern, where a broad context needs to be considered. We are currently working on linking the database with similar databases for other languages as well as with the database of historical pieces of art.[12]

At present, all Old High German texts from the period of time 700 to 1050 have been analyzed. The HiFoS data corpus consists of ca. 30 250 fully annotated single formulaic patterns. Among them, ca. 9 494 entries come from Old High German texts (ca. 700-1050), ca. 11 644 from Middle High German (ca. 1050-1350) and ca. 8 973 from Early Modern High German texts (ca. 1350-1650).

5. Ubiquitous Knowledge Processing Laboratory (UKP Lab)

Despite the many differences between historical linguistics and corpus/computer linguistics in handling formulaic patterns both disciplines show a growing interest in this phenomenon and a stronger awareness of the bilateral benefit for each other. More interdisciplinary projects need to be put forth in order to give research on formulaic patterns the solid frame that they merit with regard to the constitutive role they play in human communication, both in the past and today. The cooperation planned between the HiFoS Research Group and the UKP Lab at the Technical University Darmstadt is a good start on the way of improving the bad reputation of formulaic patterns as “a pain in the neck” of corpus and computer linguistics.

The texts from the various stages of German language need to be normalized to a common language level in order to be prepared for automatic processing. Such normalization should support the application of tools that are primarily trained on modern language, and thus may constitute a moderniza-

[12] Cf. footnote 8.

tion as described by Bollmann/Dipper/Krasselt/Petran (2012). That is, words that are no longer part of the modern language vocabulary are substituted by modern words of a similar meaning.

The Darmstadt Knowledge Processing Repository (DKPro)[13] maintained by the UKP Lab already covers many NLP components for modern languages. This includes components to handle and keep track of the modification of textual data as it is done when normalizing the texts (Eckart de Castilho/Gurevych 2009). Thus, any analysis results produced by applying the NLP components to the modernized forms can be mapped back to the original historic form. Concrete support for state-of-the-art normalization of historic texts is planned to be added to DKPro as part of the cooperation.

To aid in the identification of formulaic patterns, further components shall be added to DKPro which use afore mentioned approaches of association measures, formal preferences and distribution semantics.

We understand that the nature of historic texts and of the task at hand requires that any results of automatic processing need to be inspectable and correctable by domain experts and such corrections need to be fed back to the automatic analysis system in order improve subsequent analysis runs. To facilitate this, a task-oriented user interface needs to be implemented which conveniently exposes the functionality of the automatic analysis without bothering the user with technical details. An example how we realized such an user interface for the task of finding uncommon and ambiguous grammatical structures in large corpora has been shown in Eckart de Castilho/Bartsch/Gurevych (2012).

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant № I/82806, by the Hessian research excellence program “Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz” (LOEWE) as part of the research center “Digital Humanities”, by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1110D (DARIAH-DE), and the Alexander von Humboldt Foundation in the framework of Sofja Kovalevskaja Award 2006 (Research Group HiFoS).

[13] http://www.ukp.tu-darmstadt.de/research/current-projects/dkpro/

References

Bennett, Paul/Durrell, Martin/Scheible, Silke/Witt, Richard J. (2010): Annotating a historical corpus of German: A case study. In: Proceedings of the Conference "Language Resource and Language Technology Standards – State of the Art, Emerging Needs, and Future Developments (LREC10-W4)", 18th of May 2010: 64-68.

Bollmann, Marcel/Dipper, Stefanie/Krasselt, Julia/Petran, Florian (2012): Manual and semi-automatic normalization of historical spelling – case studies from Early New High German. In: Proceedings of the First International Workshop on Language Technology for Historical Text(s) (LThist2012), KONVENS, Vienna.

Büchler, Marco/Crane, Gregory/Mueller, Martin/Burns, Philip/Heyer, Gerhard (2011): One step closer to paraphrase detection on historical texts: About quality of text re-use techniques and the ability to learn paradigmatic relations. Chicago, IL, USA.

Burger, Harald (^{4}2010): Phraseologie. Eine Einführung am Beispiel des Deutschen. Berlin: Erich Schmidt Verlag.

Claridge, Claudia (2008): Historical corpora. In: Lüdeling, Anke/Kytö, Merja (eds.): Corpus linguistics. An international handbook. Two volumes. Berlin/New York: deGruyter, Vol. 1, 242-259.

Dipper, Stefanie (2010): POS-Tagging of historical language data: First experiments. In: Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10), Saarbrücken.

Dobrovol'skij, Dmitrij/Piirainen, Elisabeth (2005): Figurative language. Cross-cultural and cross-linguistic perspectives. Amsterdam/Philadelphia: Elsevier.

Eckart de Castilho, Richard/Gurevych , Iryna (2009): DKPro-UGD: A flexible data-cleansing approach to processing user-generated discourse. In: Online-proceedings of the first French-speaking meeting around the framework Apache UIMA at 10th Libre Software Meeting (LSM/RMLL), Nantes, France. [http://e.nicolas.hernandez.free.fr/pub/rec/09/RMLL-cfp-en.html].

Eckart de Castilho, Richard/Bartsch, Sabine/Gurevych, Iryna (2012): CSniper (2012) – Annotation-by-query for non-canonical constructions in large corpora. In: Association for Computational Linguistics: Proceedings of the 50th Meeting of the Association for Computational Linguistics (ACL) 2012 (Demo section): 85-90. [http://www.aclweb.org/anthology/P12-3015].

Evert, Stefan (2005): The statistics of word cooccurrences. Word pairs and collocations.
[http://elib.uni-stuttgart.de/opus/volltexte/2005/2371].

Evert, Stefan (2008): Corpora and collocations. In: Lüdeling, Anke/Kytö, Merja (eds.): Corpus linguistics. An international handbook. Two volumes. Berlin/New York: deGruyter, vol. 2, 1212-1249.

Filatkina, Natalia (2012): *Wan wer beschreibt der welte stat / der muoß wol sagen wie es gat.* Manifestation, functions, and dynamics of formulaic patterns in Thomas Murner's “Schelmenzunft” revisited. In: Filatkina, Natalia/Kleine-Engel, Ane/Dräger, Marcel/Burger, Harald (eds.): Aspekte der historischen Phraseologie und Phraseographie. Heidelberg: Universitätsverlag Winter, 21-41.

Filatkina, Natalia (2009a): Historical phraseology of German: regional and global. In: Korhonen, Jarmo/Mieder, Wolfgang/Piirainen, Elisabeth/Pinel, Rosa (eds.): Phraseologie global – areal – regional. Akten der Konferenz EUROPHRAS 2008 vom 13.-16.8.2008 in Helsinki. Tübingen: Niemeyer, 143-151.

Filatkina, Natalia (2009b): Historische formelhafte Sprache als „harte Nuss“ der Korpus- und Computerlinguistik. Ihre Annotation und Analyse im HiFoS-Projekt. In: Linguistik online, 39/3: 75-95.

Filatkina, Natalia/Gottwald, Johannes/Hanauska, Monika et al (2009): Formelhafte Sprache im schulischen Unterricht im Frühen Mittelalter: Am Beispiel der so genannten „Sprichwörter“ in den Schriften Notkers des Deutschen von St. Gallen. In: Sprachwissenschaft 34, 341-397.

Fellbaum, Christiane (ed.) (2006): Corpus-based studies of German idioms and light verbs. Special Issue 19-4 of the International Journal of Lexicography.

Fellbaum, Christiane (2007): Idioms and collocations. Corpus based linguistic and lexicographic studies. London: Continuum International Publisher.

Heid, Ulrich (2008): Computational phraseology. An overview. In: Granger, Sylviane/Meunier, Fanny (eds.): Phraseology. An interdisciplinary perspective. Amsterdam: John Benjamins Publishing Company, 337-360.

Heyer, Gerhard/Quasthoff, Uwe/Witting, Thomas (2008): Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse. Bochum: W3L GmbH.

Hoey, Michael (2005): Lexical priming. A new theory of words and language. London/New York: Routledge Chapman & Hall.

Jackendoff, Ray (1995): The boundaries of the lexicon. In: Everaert, Michael et al. (eds.): Idioms: Structural and psychological perspectives. New York, 133-165.

Li, Linlin/Sporleder, Caroline (2010): Using Gaussian mixture models to detect figurative language in context. In: Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010). Short Papers, June 1-6, 2010, Los Angeles. [http://www.coli.uni-saarland.de/~csporled/papers/naacl10.pdf].

Li, Linlin/Roth, Benjamin/Sporleder, Caroline (2010): Topic models for word sense disambiguation and token-based idiom detection. In: Proceedings of the 48th Annual meeting of the Association for Computational Linguistics (ACL 2010), July 11-16, 2010, Uppsala, Sweden. [http://www.coli.uni-saarland.de/~csporled/papers/acl10.pdf].

Piirainen, Elisabeth (2012): Widespread idioms in Europe and beyond. Toward a lexicon of common figurative units. New York: Peter Lang.

Rayson, Paul/Piao, Scott/Sharoff, Serge/Evert, Stefan/Moirón, Begoña Villada (2010). Multiword expressions: hard going or plain sailing? In: Language Resources and Evaluation, 44 (1): 1-5.

Sag, Ivan A. et al. (2001): Multiword expressions: A pain in the neck for NLP. In: LinGO 2001-2003. [http://lingo.stanford.edu/pubs/WP-2001-03.pdf].

Sinclair, John (1987): Collocation: A progress report. In: Steele, Ross/Threadgold, Terry (eds.): Language topics: Essays in honour of Michael Halliday. Amsterdam: John Benjamins Publishing Company, 319-331.

Sirajzade, Joshgun (2012): Das luxemburgischsprachige Œuvre von Michael Rodange (1827-1876). Editionsphilologische und korpuslinguistische Analyse. PhD Thesis, University of Trier.

Tschichold, Cornelia (2000): Multi-word units in Natural Language Processing. Hildesheim: Olms Verlag.

Wray, Alison (2008): Formulaic language: Pushing the boundaries. Oxford: Oxford University Press.

Contact

Prof. Dr. Claudine MoulinProf.
University of Trier
Trier Centre for
Digital Humanities
moulin@uni-trier.de

Dr. Iryna Gurevych
Technical University of Darmstadt
Ubiquitous Knowledge Processing (UKP) Lab
gurevych@ukp.informatik.tu-darmstadt.de

Dr. Natalia Filatkina
University of Trier
HiFoS Research Group
filatkina@uni-trier.de

Richard Eckart de Castilho
Technical University of Darmstadt
Ubiquitous Knowledge Processing (UKP) Lab
eckart@ukp.informatik.tu-darmstadt.de