Constituents in multiword expressions: What is their role, and why do we care?

Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Eva Smolka
Fachbereich Linguistik, Universität Konstanz

1 Introduction

The processing and representation of multiword expressions (MWEs), ranging from noun compounds (such as nickname in English and Ohrwurm in German) to complex verbs (such as give up in English and aufgeben in German) and idiomatic expressions (such as break the ice in English and das Eis brechen in German) have remained an unsettled issue over the past 20+ years.

Our research question concerns semantically transparent MWEs as well as MWEs that result in a meaning shift. For example, in the absence of situational experience, even complex verbs that appear to be fully semantically transparent such as aufstehen (‘stand up’) do not necessarily have whole-word meanings that are easily predictable from their constituents. Even more difficult are complex verbs such as verstehen (‘understand’) and zustehen (‘legally due’), which contain only a remote resemblance to the meaning of stehen (‘stand’). Similarly, the constituents of noun compounds do not necessarily contribute to their whole-word meanings in a straightforward way. The meaning contribution may range from relatively semantically transparent as in Nudelsuppe (‘noodle soup’) to semantically opaque, as in Spitzname (‘nickname’, lit. ‘pointy name’), Geduldsfaden (‘patience’, lit. ‘patience thread’), or Zwickmühle (‘dilemma’, lit. ‘pinch mill’), which contain a modifier (i.e. the left constituent) and/or a head (i.e. the right
constituent) that render the compound semantically more opaque. The most extreme meaning shifts across types of MWEs occur in idiomatic constructions, such as *kick the bucket* and *reach for the stars*, where the literal meanings of the constituents do not seem to contribute to the overall figurative meanings ‘die’ and ‘strive for something unachievable’ at all. MWEs of the idiomatic type are typically assumed to be semantically opaque, even though some idioms like *spill the beans* are stronger in reflecting the figurative meaning (‘reveal a secret’) in a metaphoric way than others.

This edited volume exploits complementary evidence across different types of MWEs to shed light on the interaction of constituent properties and meanings of MWEs. Specialists across languages and across research disciplines contribute to this issue and provide a cross-linguistic perspective integrating linguistic, psycholinguistic, corpus-based and computational studies.

2 Contributions

In the following, the seven contributions in this volume discuss multiword expressions that are composed of different types of constituents, including the combination of particle+stem in complex verbs (e.g., *aufstehen* ‘stand up’), the combination of stem+stem in existing and novel compounds (e.g., *nickname* and *campeel*, respectively), the combination of stem+stem+suffix in deverbal compounds (e.g., *budget assessment*), the combination of stem+preposition+stem in noun compounds (e.g., *juego de niños*), the combination of modifier+stem in modifier-noun phrases (e.g., *the brown dog*) and idiomatic combinations of words (e.g., *reach for the stars*).

Sections 2.1 to 2.3 discuss the interdisciplinary perspectives separately for complex verbs, noun compounds and idiomatic expressions, and for each of these three categories of MWEs we summarise the contributions to this collection.

2.1 Complex verbs

Seminal psycholinguistic studies have applied manipulations of semantic transparency to study whether verbal MWEs of the type prefix+stem, particle+stem and stem+suffix are lexically represented and processed via the constituents or as a whole-word unit (e.g., Taft & Forster 1975; Marslen-Wilson et al. 1994; Longtin et al. 2003).

Recurrent findings in English and French showed that semantically transparent words facilitate their base (e.g., *distrust–trust, confessor–confess*). This facilitation effect, however, was not obtained for semantically opaque primes (e.g.,
Constituents in multiword expressions

retreat–treat, successor–success). Lexicon-based models concluded from these findings that a semantically transparent word like confessor possesses a lexical entry that corresponds to its base and is represented as the stem (-confess-) and suffix (-or), whereas successor is represented in its full form (e.g., Rastle et al. 2000; Feldman et al. 2004; Diependaele et al. 2005; 2009; Meunier & Longtin 2007; Marslen-Wilson et al. 2008; Taft & Nguyen-Hoan 2010).

Semantic transparency effects emerge also when transparency is manipulated in a more graded way (Gonnerman et al. 2007): Strong facilitation effects showed for strongly phonologically and semantically related word pairs (e.g., preheat–heat), intermediate effects for moderately similar pairs (e.g., midstream–stream), and no priming for low semantically related word pairs (rehearse–hearse). Within learning-based approaches, such as the convergence-of-codes account, form and meaning relatedness between word pairs determines lexical processing (Plaut & Gonnerman 2000; Gonnerman et al. 2007).

Findings in German, however, indicate that lexical processing occurs via the stem and irrespective of semantic transparency (i.e., meaning composition of the complex verb). Low semantically related word pairs (entwerfen–werfen ‘design’–‘throw’) induced facilitation of the stem to the same extent as semantically related word pairs did: bewerfen–werfen (‘throw at’–‘throw’) (e.g., Smolka et al. 2009; 2014; 2015; 2019). Most importantly, these findings stress the importance of cross-language comparisons: what is true for the processing in one language is not necessarily true for the processing in another language (Günther et al. 2018).

Computational approaches regarding the meanings of complex verbs have mainly focused on predicting the degree of transparency of complex verbs. These approaches typically rely on the distributional hypothesis (Harris 1954; Firth 1957) and empirical co-occurrence information from large corpora, and are realised as vector space models (Turney & Pantel 2010). Regarding English, computational approaches explored variants of distributional models and distributional similarity, comparing word-based and syntax-based descriptions, large-scale vs. dimensionality-reduced representations, and verb-specific vs. general information (Baldwin et al. 2003; McCarthy et al. 2003; Bannard 2005; Cook & Stevenson 2006; i.a.). Regarding German, an initial series of papers (Aldinger 2004; Schulte im Walde 2004; 2005; 2006) studied particle verbs from a large-scale corpus-based perspective, with an emphasis on salient distributional features at the syntax-semantics interface. Schulte im Walde (2006) and Bott & Schulte im Walde (2018) integrated the subcategorisation transfer of German particle verbs with respect to their base verbs into models of compositionality. Kühner & Schulte im Walde (2010), Bott & Schulte im Walde (2017), and Köper & Schulte im Walde
Sabine Schulte im Walde & Eva Smolka

(2017a) used clustering to distinguish between multiple senses, and common cluster membership to determine compositionality. Köper & Schulte im Walde (2016) and Aedmaa et al. (2018) applied classifiers to identify figurative language usage of German and Estonian particle verbs in context.

So far, most approaches that have dealt with complex verbs – across disciplines and across languages – have considered semantic transparency as the meaning relation between the whole word meaning of the MWE and the meaning of its base constituent, disregarding the contribution of the often ambiguous prefix or particle, e.g., they were concerned with the question: to what degree is the meaning of stand reflected in understand? Apart from a series of formal word-syntactic analyses in the framework of Discourse Representation Theory (Kamp & Reyle 1993) for German particle verbs with the particles auf (Lechler & Roßdeutscher 2009), ab (Kliche 2011), nach (Haselbach 2011) and an (Springorum 2011), this gap of knowledge has recently been addressed from experimental perspectives: Frassinelli et al. (2017) demonstrated in a lexical decision experiment that the particle an in German particle verbs is primarily associated with a horizontal directionality, while auf is primarily associated with a vertical directionality. Schulte im Walde et al. (2018) and Köper & Schulte im Walde (2018) present data collections to assess meaning components in German complex verbs. The former dataset contains source- and target-domain characteristics of the base verbs and the complex verbs, respectively, and a selection of arrows to add spatial directional information to user-generated contexts; the latter dataset contains ratings for strengths of particle-related pairs of German base verbs and particle verbs.

As part of the present collection, Springorum & Schulte im Walde also focus on the meaning contribution of the particle to the overall meaning of German particle verbs. They combine nine particles (e.g., auf ‘up’) with 30 base verbs (e.g., geben ‘give’) and examine how the particles are perceived in adding directionality (i.e., up, down, left, right) to the meaning of the particle verb (e.g., aufgeben ‘give up’). That is, the participants in their study saw a base verb or a particle verb and decided which type of directionality in form of two-dimensional arrows best reflects the verbal meaning. Their qualitative and quantitative analyses indicate that the particles exhibit individual spatial profiles, but also that the particles vary in their flexibility to provide predominant directions, in interaction with the abstractness of the semantic base verb domains.

2.2 Noun compounds

Compounds also lie on a continuum between relatively transparent and rather opaque with respect to the meanings of their constituents. Psycholinguistic re-
search so far has been intrigued by the question whether the compound is lexically represented and processed via the constituents or as a whole-word unit. For example, findings on the processing of noun-noun compounds indicate a competition between the compounds’ constituents that correspond to independent words and their whole-word counterparts. Hence, upon seeing the compound *doughnut*, the constituent 
[nut] may compete with the whole word *nut* (e.g., Libben 2006; Frisson et al. 2008; Monahan et al. 2008; Fiorentino & Fund-Reznicek 2009; Gagné & Spalding 2009; 2014; Libben 2014). Another question concerns whether the semantic transparency of the constituents affect the processing of the MWE they compose, and if so, how? Indeed, semantically opaque compounds are generally processed more slowly than semantically transparent ones, and are less likely to show constituent activation – probably because the semantic opacity of the whole compound makes its constituents less relevant to lexical comprehension (e.g., Taft & Forster 1975; Sandra 1994; Zwitserlood 1994; Isel et al. 2003; Libben et al. 2003). Furthermore, recent studies indicate that the influence of semantic transparency is language-specific. The semantic transparency of the head has been found to affect the processing of noun-noun compounds in English and Italian (e.g., Marelli et al. 2009; Marelli & Luzzatti 2012) but not in German (e.g., Smolka & Libben 2017).

Computational approaches to predicting the transparency of noun compounds can be subdivided into two subfields:

1. approaches that aim to predict the *meaning* of a compound by composite functions, relying on the vectors of the constituents (e.g., Mitchell & Lapata 2010; Coecke et al. 2011; Baroni et al. 2014; Hermann 2014); and

2. approaches that aim to predict the *degree of compositionality* of a compound, typically by comparing the compound vectors with the constituent vectors (e.g., Reddy et al. 2011; Salehi & Cook 2013; Schulte im Walde et al. 2013; Salehi et al. 2014a,b; 2015; Schulte im Walde et al. 2016; Köper & Schulte im Walde 2017b).

As for complex verbs, the computational models under 2. typically rely to a large extent on the distributional hypothesis and empirical co-occurrence information from large corpora. Individual research studies noticed differences in the contributions of modifier and head constituents towards the composite functions predicting compositionality (Reddy et al. 2011; Schulte im Walde et al. 2013), but only a very limited number of approaches zoomed into potentially relevant properties of MWEs and their constituents, such as ambiguity, frequency and productivity (Bell & Schäfer 2016; Schulte im Walde et al. 2016).
In this collection, Pezzelle & Marelli apply a distributional semantic model to show that the semantic properties of the compound and its constituents may explain syntactically-based classes of compounds as suggested in linguistic theories (Bisetto & Scalise 2005). They differentiate between types of compounds such as subordinate, attributive, and coordinate compounds, on the basis of the underlying syntactic relation between the compound constituents. In particular, Pezzele and Marelli provide measures that quantify (a) the degree of semantic similarity between the constituents, and (b) the contribution of each constituent to the overall compound meaning, and show that these semantic measures are effective in capturing the different syntactic linguistic classes. In other words, the continuous quantitative semantic aspects of the meanings of compounds parallel the discrete qualitative grammatical distinctions between compounds.

Iordăchioaia, van der Plas & Jagfeld study the compositionality of English deverbal compounds. These deverbal nouns are ambiguous between compositionally interpreted “argument structure nominals”, which inherit verbal structure and realise arguments (e.g., assessment of the budget by the government), and more lexicalized “result nominals”, which preserve no verbal properties or arguments (e.g., budget assessment), cf. Grimshaw (1990). While the former are fully compositional, the latter remain ambiguous because the non-head (budget) can be interpreted as either subject or object. The authors apply machine-learning techniques to evaluate corpus data and human annotations to support their hypothesis and find that different properties of the head contribute to the interpretation of the deverbal compound.

In the third chapter on compounds, Libben investigates English compounds from a psycholinguistic perspective. He uses novel compounds such as ankle-cob and clampeel, the former being unambiguous, the latter being ambiguous in the way they can be parsed (i.e. ankle-cob versus clam-peel or clamp-eel, respectively). A typing experiment shows that the typing latencies indeed peak at the morpheme boundary of non-ambiguous compounds. Equivalent latencies at the critical letters of ambiguous compounds indicate that they are parsed in both possible reading ways. Libben refers to the heuristics of his Fuzzy Forward Lexical Activation account, which assumes that MWEs are parsed from left to right for any possible word combination. He concludes that complex words are not static representations but rather patterns of actions.

Two papers deal with MWEs that are untypical compound constructions for which linguistic theories in general refer to the notions of lexicon and syntax and debate whether these MWEs are to be considered as compounds or not. Hennecke examines the formation of MWEs of the type “N Prep N” in Romance languages, such as Spanish, French and Portuguese (e.g. juego de niños, ‘kid’s game’).
and takes a constructionist approach to analyse the constructions as abstract templates. In a qualitative analysis, she examines the variation that the preposition in a construction may undergo (e.g. juego de niños vs. juego para niños, both meaning 'kid’s game'). To this end, she analyses the semantic relations between the nominal constituents and the semantic transparency of the constructions. Her findings indicate that variability of the prepositional element occurs only in semantically transparent constructions. Furthermore, prepositional variability largely varies across the three Romance languages.

Also Gagné, Spalding, Burry & Adams examine MWEs that are not typically classified as compounds and compare modifier-noun phrases (e.g., the brown dog) with full phrases (e.g., the dog that was brown). They examine how modifying information that refers to recently encountered information is used in the production of MWEs, and manipulate the property of the head noun between normal (e.g., brown) and distinctive (e.g., blue). Participants showed a strong overall bias toward using a modifier-noun phrase structure (regardless of whether they previously saw a modifier-noun phrase or a full phrase), and were more likely to include distinctive properties (the blue dog) than normal properties (the brown dog) when referring to the concept. These findings indicate that modifier-noun phrases have a privileged status among MWEs and provide a good compromise between conveying sufficient information and using simple syntactic structures.

2.3 Idioms

Idiomatic expressions are the MWEs which may be considered as showing the strongest semantic shift that the constituents undergo, because the figurative meaning is usually not even remotely connected with the meaning of its constituents, as in hit the road. Rather, idiomatic expressions are considered semantically fixed, since the figurative meaning does not allow the replacement of any of the word constituents (e.g., *she hit the street; *she beat the road), and the modification of an idiomatic constituent is assumed to change the figurative meaning into a literal meaning.

The processing and representation of idioms has thus remained an unsettled issue in psycholinguistic research: how is the figurative meaning processed and stored in lexical memory? In particular, is the figurative meaning of an idiom represented separately from the meaning of its constituents, and how is the figurative meaning assembled (e.g., Cacciari & Tabossi 1988; Gibbs Jr. 1992; Cacciari & Glucksberg 1994; Titone & Connine 1999; Hamblin & Gibbs Jr. 2003)? Seminal studies thus assumed a “non-compositional” representation in which the whole figurative meaning of an idiom is stored as a distinct entry in the mental lexicon.
similar to the representation of a complex word like *Finanzmarktaufsichtsbehörde* (‘financial market supervisory authority’) (e.g., Bobrow & Bell 1973; Swinney & Cutler 1979; Gibbs Jr. 1980). More recent hybrid models try to integrate the assumption that idioms are both compositional and unitary: on the one hand, an idiom is composed of single constituents that are activated to some degree, and on the other hand each idiom possesses its own lexical entry that stores the whole meaning of the idiom (e.g., Cacciari & Tabossi 1988; Gibbs Jr. et al. 1992; Cutting & Bock 1997; Titone & Connine 1999; Sprenger et al. 2006; Caillies & Butcher 2007; Holsinger & Kaiser 2013; Titone & Libben 2014).

As far as computational work on idiomatic expressions is concerned, several research studies measured the syntactic flexibility of idiomatic expressions, to a large extent focusing on verb–object combinations (e.g., Bannard 2007; Fazly et al. 2009). These measures varied the constituents of the target MWEs, explored modifiability and passivisation, etc. in order to distinguish between literal vs. idiomatic interpretations. A large number of automatic classification approaches addressed idioms as non-literal language across various types of MWEs, mostly relying on contextual indicators to distinguish between literal and idiomatic interpretations (e.g., Sporleder & Li 2009; Turney et al. 2011; Köper & Schulte im Walde 2016), such as distributional similarity, text cohesion graphs, and contextual abstractness. The variation-based approaches further provide some insight into the flexibility of the constituents of MWEs and their meaning contributions.

The last paper by Smolka & Eulitz deals with idioms and how the meaning of the constituents contributes to the figurative meaning. They present three experiments, in which participants rate the meaning similarity between an idiomatic phrase (e.g., *She always reached for the stars*) and a paraphrase of its figurative meaning (e.g., *She always strove for something unreachable*). They exchange the noun, verb, or prepositional idiomatic constituent by a close semantic associate (e.g., *She always reached/grasped for/at the stars/planets*) and find that a modified constituent still preserves the figurative meaning. This study adds to the understanding that there is no completely fixed unitary entry and that the idiomatic constituents do contribute to the figurative meaning of the idiom, even though the figurative meaning is semantically opaque.

**Acknowledgements**

This collection was supported by the DFG Collaborative Research Centre SFB 732 and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde), and by the Volkswagen Foundation Grant FP 561/11 (Eva Smolka). Special thanks
go to our student researcher Anurag Nigam who type-set this volume. Last but not least we thank our experts from the interdisciplinary fields who ensured a qualitatively high-standing reviewing process:

- Melanie Bell (Anglia Ruskin University, UK)
- Jens Bölte (University of Münster, Germany)
- Paul Cook (University of New Brunswick, Canada)
- Christina Gagné (University of Alberta, Canada)
- Giannina Iordăchioaia (University of Stuttgart, Germany)
- Alessandro Lenci (University of Pisa, Italy)
- Marco Marelli (University of Milano-Bicocca, Italy)
- Carlos Ramisch (Aix-Marseille University, France)
- Martin Schäfer (University of Jena, Germany)
- Nils Schiller (Leiden University, The Netherlands)
- Thomas Spalding (University of Alberta, Canada)
- Lonneke van der Plas (University of Malta, Malta)
- Aline Villavicencio (University of Sheffield, UK / Federal University of Rio Grande do Sul, Brazil)

References


Constituents in multiword expressions


1 Constituents in multiword expressions


Sabine Schulte im Walde & Eva Smolka


1 Constituents in multiword expressions


Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1). 141–188.
1 Constituents in multiword expressions