German Synthetic Compounds and the Architecture of the Grammar:
A Behavioral Analysis

The aim of this paper is to consider the place of German synthetic compounds (hence SCs, in German *Rektionskomposita*; cf. Gaeta 2010 for a recent discussion) within the general architecture of the grammar, using new empirical methods which have become available thanks to the advent of very large electronic corpora. In particular, we will pose the question: is it possible to predict the lexical selectional behavior of SCs from data on the usage of their related verbal head lexemes? By focusing directly on attested heads and modifiers, we wish to shed light on how we can account for the structure of SCs. Two main hypotheses can be formulated here which also presuppose a different architecture of the grammar: a syntactic hypothesis, wherein SCs are formed on the basis of syntactic inputs which are elaborated within a lexical component (cf. theoretical frames pleading for syntax below zero such as Distributed Morphology, see Spencer 2005 for an overview); and a lexical hypothesis, where SCs result from elaboration of lexical patterns represented in the lexical component (lexicalist approaches to morphology, cf. Scalise & Guevara 2005 for a survey).

In order to assess each hypothesis, we will proceed empirically. Working within a usage-based approach, we attribute relevance to the actual attested spectrum of arguments observed with a lexeme realized as a verb and/or an SC head. If the syntactic hypothesis is correct, then we expect a substantial correspondence between attested syntactic input and the output of SCs, which should form a subset thereof. On the other hand, if this proves to be false, we expect, on the basis of the lexical hypothesis, that lexical patterns occur which account for those SCs which do not exhibit a syntactic base, at least for infrequent, novel compounds which are not expected to be lexicalized.

Our investigation is based on German compounds headed by deverbal nouns formed with the suffix *-er*, e.g. *Mäusefänger* 'mouse catcher', *Romanleser* 'novel reader', etc. This suffix generally forms agent nouns, although it can also give rise to instruments and to several other minor meanings (cf. Meibauer et al. 2004). We utilize a large web corpus consisting of approximately 1.63 billion word forms (DeWaC, see Baroni et al. 2009) in order to automatically extract SCs headed by nouns suffixed with *-er* and corresponding verbs and direct objects, filtered using a high precision rule based procedure. To the best of our knowledge, our investigation is the most extensive empirical analysis of German SCs to date.

Our results are incongruent with the syntactic hypothesis, at least in its strongest form. As it turns out, many lexemes show a preference for, and wider variety of compound or verbal types. We can identify thousands of combinations attested only in one construction, forming several interesting groups. These include lexicalized compounds such as *Krankheitserreger* 'pathogen, (lit.) disease exciter' and verbal collocations such as *Gedanken machen* '(lit.) make thoughts, think' (cf. *Krankheit erregen* 'excite disease' and *Gedankenmacher* 'thought maker' which are not absent in our data); but also productive combinations, e.g. professions headed by lexicalized agent nouns like *Leiter* 'leader, head', *Hersteller* 'manufacturer', *Sammler* 'collector' and many others, as well as ad-hoc attributive formations such as *Ausbildungsplatzbesitzer* 'apprenticeship place possessor'. In some cases the relevant lexeme is 'defective', because a head avoids *-er* nominalization (*Kinder bekommen* 'get kids' but no ?*Kinderbekommer* 'kid getter'), or the corresponding verb form is obsolete, though the lexeme remains productive in compounding, e.g. *Staubsaugervertreter* 'vacuum cleaner salesman' but synchronically absent *Staubsauger vertreten* '(lit.) to represent vacuum cleaners'. Such compounds are most straightforwardly accounted for on the basis of frequent lexical patterns extended with the help of a (prototype-based) analogical mechanism.

On the other hand, we also find evidence that the contribution of possible syntactic inputs is very large for many lexemes with a high lexical correspondence, which, at least indirectly, calls into play a mechanism of reinforcement of the lexical pattern with the concurrent action of the syntactic component, however it may be conceived of. Thus, we concur with a dual model of the kind suggested by Schreuder & Baayen (1995) to deal with the lexical processing of inflected words, which seems appealing in order to explain both the extensive productivity of syntax-based SCs and the lexically based patterns which resist a syntactic analysis.

References

Baroni, Marco, Bernardini, Silvia, Ferraresi, Adriano & Zanchetta, Eros (2009), "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora". *Language Resources and Evaluation* 43.3: 209-226.

Gaeta, Livio (2010), "Synthetic Compounds. With Special Reference to German". In Sergio Scalise and Irene Vogel (eds.), *Cross-Disciplinary Issues in Compounding*. Amsterdam & Philadelphia: John Benjamins, 219-235.

Meibauer, Jörg, Anja Guttropf & Carmen Scherer (2004), "Dynamic Aspects of German *-er*-Nominals: A Probe into the Interrelation of Language Change and Language Acquisition". *Linguistics* 42.1: 155-193.

Scalise, Sergio & Emiliano Guevara (2005), "The Lexicalist Approach to Word-formation and the Notion of the Lexicon". In Pavol Štekauer and Rochelle Lieber (eds.), *The Handbook of Word-Formation*. Dordrecht: Springer, 147-187.

Schreuder, Robert & Harald Baayen (1995), "Modeling Morphological Processing". In Laurie B. Feldman (ed.), *Morphological Aspects of Language Processing*. Hillsdale, NJ: Erlbaum, 131-154.

Spencer, Andrew (2005) "Word-formation and Syntax". In Pavol Štekauer and Rochelle Lieber (eds.), *The Handbook of Word-Formation*. Dordrecht: Springer, 73-97.