Published in: Corpas Pastor, Gloria/Monti, Johanna/Seretan, Violeta/Mitkov, Ruslan (ed.): EUROPHRAS 2015, 29 June - 2 July 2015, Malaga, Spain. Proceedings of the workshop on multi-word units in machine translation and translation technology (MUMTTT 2015), 1-2 July. Geneva: Editions Tradulex, 2016, pp. 13 - 18.

Corpus-driven Description of Multi-word Patterns

Kathrin Steyer
Institut für Deutsche Sprache
Mannheim
steyer@ids-mannheim.de

Abstract

This paper presents our model of 'Multi-Word Patterns' (MWPs). MWPs are defined as recurrent frozen schemes with fixed lexical components and productive slots that have a holistic – but not necessarily idiomatic – meaning and/or function, sometimes only on an abstract level. These patterns can only be reconstructed with corpus-driven, iterative (qualitative-quantitative) methods. This methodology includes complex phrase searches, collocation analysis that not only detects significant word pairs, but also significant syntagmatic cotext patterns and slot analysis with our UWV Tool. This tool allows us to bundle KWICs in order to detect the nature of lexical fillers for and to visualize MWP hierarchies.

First we discuss the nature of MWPs as frozen communicative units. Then, we illustrate our methodology and selected linguistic results using examples from a contrastive study of German, Spanish, and English prepositional MWPs.

1 Introduction

Learning from corpora does not just mean to find a certain number of similar citations that confirm a hypothesis. It means knowledge about patterns of language use. Patterns can be reconstructed from corpus analysis by collecting many similar use cases – bottom up in a corpus-driven way. Looking at many use cases does not mean describing what is already known and visible: It means seeing hidden structures. This is not merely 'more data', but new interrelations, unusual cross-connections, surprising relationships, and networks. Of course, pattern detection is not a new invention but one of the central methods in information science, data mining, and information retrieval. But, we are convinced that in respect to a qualitative reconstruction of hidden patterns in language use and their applications in lexicography and second language teaching, we are just at the beginning. We would like to discuss this pattern view of language use on the basis of multi-word expressions and phrasemes.

2 Multi-word Patterns

Due to the rise of corpus linguistics and the feasibility of studying language data in new quantitative dimensions, it became more and more evident that language use is fundamentally made up by fixed lexical chunks, set phrases, long-distance word groups, and multi-word expressions (MWEs). Sinclair's inductively reconstructed collocations (cf. 1991) and Hausmann's collocation pairs (cf. 2004) are the two leading concepts in collocation research. Basically, they are merely different ways of looking at the same fundamental principle of language: linguistic frozenness and fixedness. Compositional collocations and idioms differ in their degree of lexical fixedness and semantic opacity, their recognisability and prototypicality (cf. Moon 1998, Burger et al. 2007). But they all share the most important characteristic: They are congealed into autonomous units in order to fill a specific role in communication. All these fragments are fixed patterns of language use (cf. Hunston/Francis 2000; cf. Hanks 2013). There is no core and no periphery. The difference is only in the degree of conspicuousness for the observer. These word clusters did not become fixed expressions by chance, but because there was a need of speakers for an economic way of communicating (cf. Steyer 2013). Currently, this widening of scope to every kind of frozen multi-word unit is also accepted in modern phraseology, as Dobrovol'skij outlined in 2011 in the third volume of "Konstruktionsgrammatik" in a very compact way.

Lately, not only multi-word research but also usage-based linguistics as a whole is subject to a shift. If you conduct empirical studies on corpora systematically and – this is very important – in a bottom up way, it is evident that MWEs are not as singular and unique as it is often still assumed in phraseology. MWEs are linked in many ways with other units in the lexicon. They are specific lexical realisations of templates, definitely more noticeable and more fixed than ad-hoc formulations, but not unique. Such templates emerge from repeated usage and can be filled with ever changing lexical elements, both phraseological and non-phraseological. We call them 'Multi-word Patterns' (MWPs) (cf. Steyer 2013)[1].

MWPs are recurrent frozen schemes with fixed lexical components and productive slots that have holistic – but not necessarily idiomatic – meanings or functions, sometimes only on an abstract level. The slots are filled with lexical units that have similar lexical-semantic and/or pragmatic characteristics, but must not belong to the same morpho-syntactic class. Speakers are able to recall those schemes as lexicon entries and fill the gaps in a specific communicative situation in a functionally adequate way. For example, the sentence *Die Worte* ***klingen fremd für westliche Ohren*** (*The words* ***sound strange for Western ears***) is based on the following MWP:

(1)
für* X *Ohren Y ***klingen***
(ww: to sound Y for X ears)

X ADJ$_{\{HUMAN\}}$ fillers: *deutsche* (German) / *westliche* (Western) / *europäische* (*European*) /…

Y ADV$\{_{CONNOTATION}\}$ fillers: *fremd* (foreign) / *ungewohnt* (unfamiliar) / *exotisch* (exotic) / *seltsam* (strange) / *vertraut* (familiar) / *merkwürdig* (odd) / *schräg* (discordant) / *pathetisch* (melodramatic) /...

Holistic Meaning:
'Somebody (a person / a group of people / a specific community) could possibly perceive, interpret, or assess something in a certain way'

The X ADJ fillers refer to a person, to groups of people, or to specific communities. The Y ADV collocations are almost always connotative adverbs. The whole pattern expresses specific interpretations of a fact or situation. But the speaker do not present the interpretation or evaluation as his own. He pretends that this is the interpretation of an abstract or fictional group of people. So the speaker can present the interpretation as possible or given without having to take responsibility for it.

MWEs and MW patterns are not clear-cut or distinct entities. On the contrary, fragments and overlapping elements with fuzzy borders are typical for real language use. This means that there are rarely MWEs as such. In real communicative situations, some components are focused while others fade to the background.

The reconstruction of MWPs is only possible with complex corpus-driven methods in an iterative way (quantitative – qualitative).[2] Generally, we study the nature of MW patterns by exploring keyword-in-context concordances (KWIC) of multi-word units. Beside complex phrase searches and reciprocal analysis with COSMAS II (cf. CII), we use mainly two empirical methods for KWIC bundling: We assess collocation profiles that are calculated by the IDS collocation analysis algorithm (cf. Belica 1995). This type of collocation analysis bundles KWICs and citations according to the LLR (log likelihood ratio) and also summarizes the results as lists of collocation clusters and syntagmatic patterns (compare Figure 1 in 3.). The second method is exploring and bundling KWICs with our UWV Tool that allows us to define search patterns with specific surface characteristics, depending on our research question or hypothesis (cf. Steyer/Brunner 2014). The search patterns are essentially regular expressions consisting of fixed lexical items and gaps between those (with an arbitrary length, i.e., the fillers do not have to be single words, but can also be n-grams). The fillers are ranked according to frequency, and it is also possible to annotate them with tags, to add narrative comment, and to output annotated filler groups. All this interpreted data can be exported for a lexicographic online representation, recently as "Multi-Word Fields" (cf. Steyer et al. 2013).

[1] This term is similar to the term 'phrasem-constructions' proposed by Dobrovols'kij in 2011. But we prefer Steyer's term because we do not want to focus on the construction grammar framework, but take a strictly lexical and first and foremost usage-based perspective. Without doubt, the discussion of the relationship between these approaches is high on our agenda.

[2] The following examples are all taken from the *German Reference Corpus* (*Deutsches Referenzkorpus*) (cf. DeReKo), located at the Institut for the German Language in Mannheim. Our focus lies on syntagmatic word surface structures, and we use corpora that are not morpho-syntactically annotated.

In the following chapter, we illustrate our methodology and selected linguistic results using examples from a new contrastive project (German – Slovakian – Spanish)[3].

We concentrate on the German - Spanish contrast (with added English examples), but the main aspects can also be observed in Slovakian.

3 MWP in Contrast – Methods and Evidences

Our research goal is the detection and description of prepositional MWE and MW patterns like *nach Belieben* (at will), *mit Genugtuung* (with satisfaction), *am Ende* (at the end). We explore and describe their fixedness, variance, and usage on several levels of abstraction and in interlingual contrast.

The key questions are:

- On which level can we find differences in the use of prepositional MWE and patterns in the three languages?
- Are there parallels on higher levels of abstraction that allow us to assume universal functional concepts?
- Is it possible to visualize these relationships and if yes, which kind of representation is appropriate for which audience, for example for foreign language acquisition?

The following two aspects are in the center of our multilingual analysis: a) collocation fields in contrast and b) lexical filler and cotext patterns in contrast. We will now look at the MWP *mit Genugtuung* (*con satisfacción / with satisfaction)* as an example.

With the help of collocation profiles calculated with CA for German and with Sketch Engine for Spanish and English (see Figure 1) we describe the meaning and usage and identify phenomena of convergence and divergence:

Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
5806	5806	103577	**Genugtuung**	100% mit Genugtuung
6762	956	9998	**Kenntnis**	99% mit Genugtuung zur Kenntnis genommen
7267	505	2930	**fest**	99% stellte mit Genugtuung fest dass
7583	316	2410	**erfüllt**	79% erfüllt [mich ...] mit Genugtuung ...
7950	367	2229	**genommen**	98% mit Genugtuung zur Kenntnis genommen
8179	229	1983	**registriert**	75% mit Genugtuung registriert
8457	278	1783	**aufgenommen**	98% mit Genugtuung aufgenommen worden
8660	203	1579	**feststellen**	99% mit Genugtuung feststellen dass ...
8885	225	1154	**nahm**	87% nahm [das ...] mit Genugtuung zur|auf

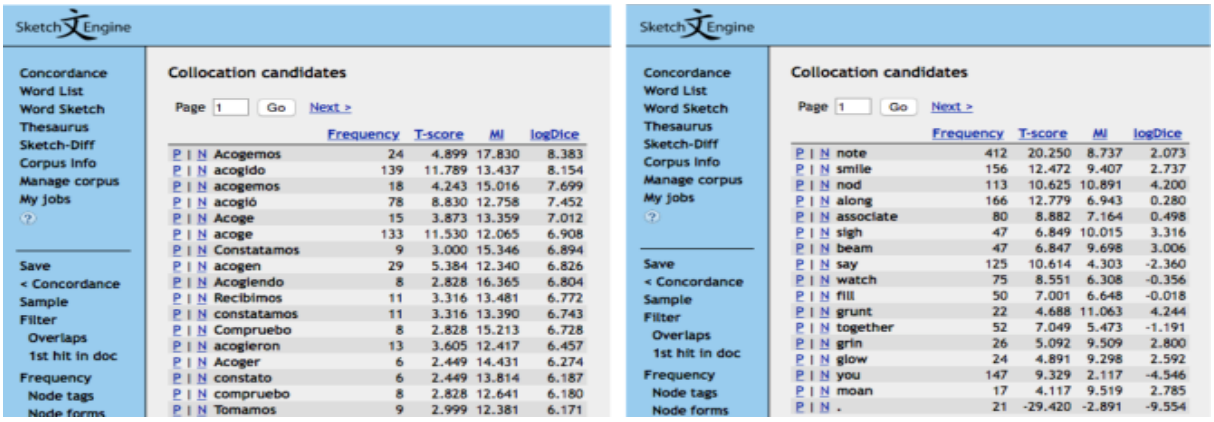


Figure 1. Collocation profiles *mit Genugtuung* (cf. CA) and *con satisfacción – with satisfaction* (cf. SKE) (snippet)

[3] Three partner institutions are involved in this research project: the University of Santiago de Compostela (Head: Carmen Mellado Blanco), the University of Trnava (Head: Peter Ďurčo), and the IDS with the UWV research group (Head: Kathrin Steyer).

The collocation profiles give us strong evidence for a restriction of verbal collocation partners: This multi-word expression is prototypically combined with verbs that refer to communicative acts:

(2)
[*mit Genugtuung* V]:
V partners: *mitteilen / sagen / hinweisen / ankündigen / zur Kenntnis nehmen /...*
[*con satisfacción* V]
V partners: *constatar* (to be stated) / *recibir* (to admire) / *saludar* (to appreciate) / *observar* (to observe) /...
[*with satisfaction* V]:
V partners: *note / say / remark / reflect /...*

Because of the verbal convergence, you can assume an interlingual abstract pattern:

[*mit / con / with* SUB{EMOTION} V{COMMUNICATION}]

An interesting difference can be observed between German, on the one hand, and Spanish and English on the other hand: Many verbal collocation partners on the highest ranks of the Spanish *con satisfacción* and the English *with satisfaction* refer to non-verbal behavior like *nod / smile / beam / grunt* resp. *reír* (to laugh) / *sonreír* (to smile) / *suspirar* (to sigh) / *respirar* (to breathe) / *fruncir los labios* (to purse one's lips). In German, this kind of contextualization is a very rare phenomenon.

In a second step, we generate filler tables with the help of our UWV Tool and compare them between the languages (see Figure 2):

First of all, the tables give information concerning the degree of lexical fixedness. As Figure 2 shows, the gap between the preposition *mit* and the noun *Genugtuung* is empty in approx. 70% of occurrences. This "Zero Gap" indicates a high degree of lexicalization and a lexicon entry *mit Genugtuung*. In Spanish and English, this empty slot is not so recurrent. Instead a strong internal variance is established.

In all three tables, we can observe groups of ADJ fillers with the same communicative functions: a) intensification, e.g., *groß – gran – great*, and b) connotation, e.g., *grimmiger – insana* (insane) – *grim*. In many cases, both functions overlap.

As mentioned in Chapter 2, the UWV tool enables to define any size of slots. Figure 3 (see next page) illustrates – for example – typical trigram fillers in German, Spanish and English.

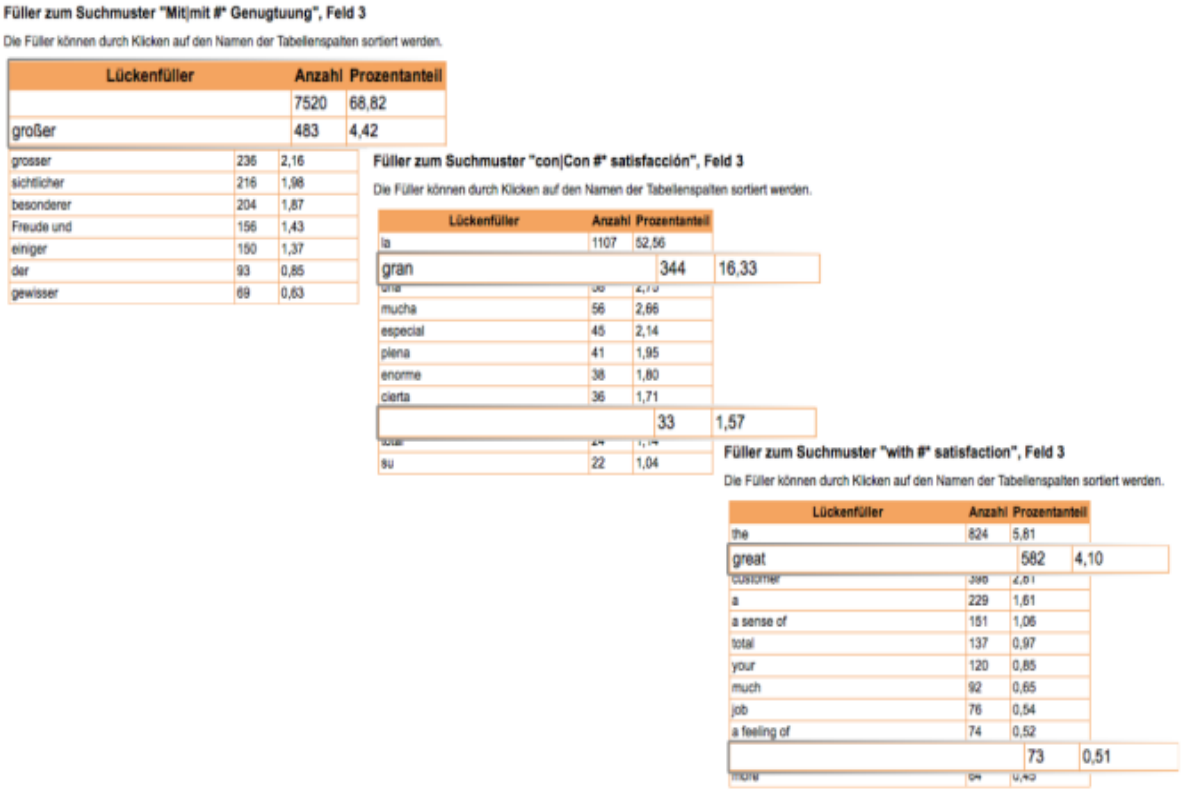
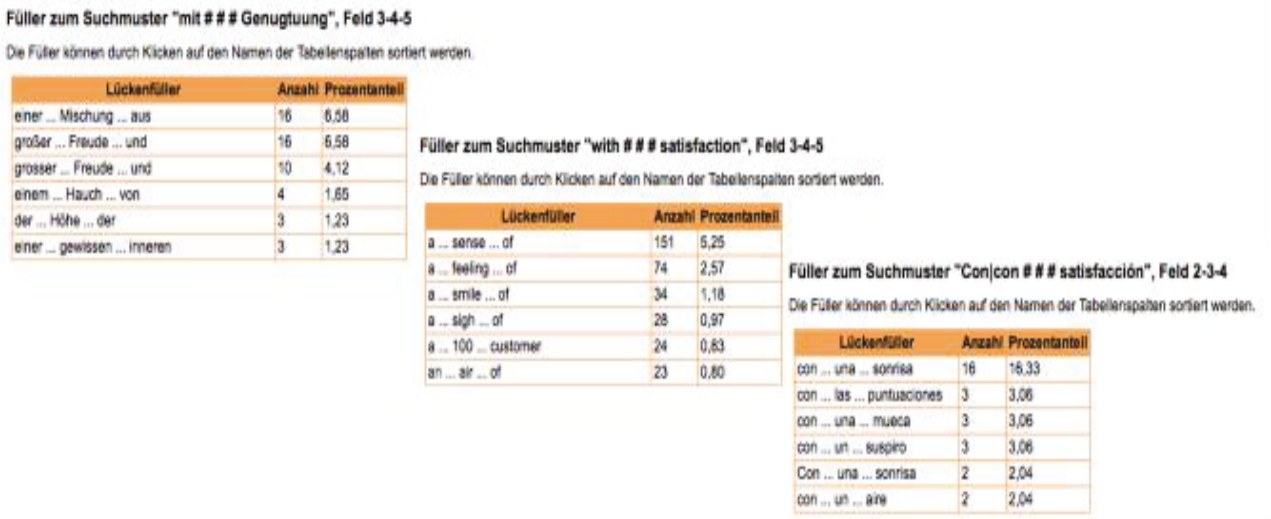


Figure 2. Filler tables of *mit – con –* with # (1 slot) *Genugtuung – satisfacción – satisfaction* (snippet)



Füller zum Suchmuster "mit # # # Genugtuung", Feld 3-4-5

Die Füller können durch Klicken auf den Namen der Tabellenspalten sortiert werden.

Lückenfüller	Anzahl	Prozentanteil
einer ... Mischung ... aus	16	6,58
großer ... Freude ... und	16	6,58
grosser ... Freude ... und	10	4,12
einem ... Hauch ... von	4	1,65
der ... Höhe ... der	3	1,23
einer ... gewissen ... inneren	3	1,23

Füller zum Suchmuster "with # # # satisfaction", Feld 3-4-5

Die Füller können durch Klicken auf den Namen der Tabellenspalten sortiert werden.

Lückenfüller	Anzahl	Prozentanteil
a ... sense ... of	151	5,25
a ... feeling ... of	74	2,57
a ... smile ... of	34	1,18
a ... sigh ... of	28	0,97
a ... 100 ... customer	24	0,83
an ... air ... of	23	0,80

Füller zum Suchmuster "Con|con # # # satisfacción", Feld 2-3-4

Die Füller können durch Klicken auf den Namen der Tabellenspalten sortiert werden.

Lückenfüller	Anzahl	Prozentanteil
con ... una ... sonrisa	16	16,33
con ... las ... puntuaciones	3	3,06
con ... una ... mueca	3	3,06
con ... un ... suspiro	3	3,06
Con ... una ... sonrisa	2	2,04
con ... un ... aire	2	2,04

Figure 3. Trigram filler tables *mit – with – con # # # Genugtuung – satisfaction – satisfacción* (snippet)

Interesting phenomena of convergence are recurrent evaluative quantifier or intensifier phrases in all three languages:

(3)
[*mit* X *Genugtuung*]
X fillers: *einem Hauch von / einem Anflug von / einem Schuss von / einer Prise von /...*
[*con* X *satisfacción*]
X fillers: *mayor nivel de* (higher level of) */ un grado de* (a degree of) */ una pizca de* (a pinch of) */ algún grado de* a degree (some degree of) */...*
[*with* X *satisfaction*]
X fillers: *a sense of / a feeling of / a great deal of /...*

This suggests an interlingual tendency to express a scale of satisfaction in a more or less indirect way.

Another example of convergent bigram fillers are coordinative structures, e.g., appositions of nouns with positive connotations.

(4)
[*mit* X ***und*** *Genugtuung*]
N fillers: *Stolz / Freude* / *Häme* (scorn)
[*con* X ***y*** *satisfacción*]
N fillers: *orgullo* / *alegría* /*asombro* (wonder)
[*with* X *and satisfaction*]
N fillers: *pride* / *joy* / *pleasure*

Because of these pronounced similarities we assume universal abstract patterns with a holistic function of intensification.

(5)
[*mit* $N_{\{EMOTION\}}$ + *und* + *Genugtuung*]
[*con* $N_{\{EMOTION\}}$ + *y* + *satisfacción*]
[*with* $N_{\{EMOTION\}}$ + *and* + *satisfaction*]

[*mit* $N_{\{EMOTION\}}$ + *und* + $N_{\{EMOTION\}}$]
[*con* $N_{\{EMOTION\}}$ + *y* + $N_{\{EMOTION\}}$]
[*with* $N_{\{EMOTION\}}$ + *and* + $N_{\{EMOTION\}}$]

[P + $N_{\{EMOTION\}}$ + *und / y / and* + $N_{\{EMOTION\}}$]

In our lexicographic description, we will try to show convergences and divergences between the languages with the aid of collocation fields and slot-filler tables on several levels of abstraction. These will be annotated and systematized according to typical usage characteristics and linked across languages.

4 Conclusion

If patterns and imitation are the genuine principles of language production and reception, they must move to the focus of lexicographic description, language acquisition, and machine translation. How these highly complex, overlapping phenomena can be structured and explained in a didactically effective way will be one of the most exciting questions for future researches in these fields.

Reference

Belica, Cyril 1995. *Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode.* Institut für Deutsche Sprache, Mannheim.

Burger, Harald, Dmitrij Dobrovol'skij, Peter Kühn and Neal R. Norrick. 2007 (eds.). *Phraseologie. Ein internationales Handbuch zeitgenössischer Forschung/Phraseology. An international Handbook of Contemporary Research.* (2 Editions). (HSK 28, 1/2). de Gruyter, Berlin/New York:.

Dobrovol'skij, Dmitrij. 2011. Phraseologie und Konstruktionsgrammatik. In Lasch, Alexander and Alexander Ziem (eds.), *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze*. (Stauffenburg Linguistik 58). Stauffenburg, Tübingen: 111-130.

Hanks, Patrick. 2013. *Lexical Analysis. Norms and Exploitations*. The MIT Press, Cambridge, MA /London.

Hausmann, Franz Josef 2004. 'Was sind eigentlich Kollokationen?' In Steyer, Kathrin (ed.), *Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003*. de Gruyter, Berlin/New York: 309-334.

Hunston, Susan and Gill Francis. 2000. *Pattern Grammar. A corpus-driven approach to the lexical grammar of English.* Amsterdam/Philadelphia: John Benjamins.

Moon, Rosamund 1998. *Fixed Expressions and idioms in English. A Corpus-Based Approach*. Clarendon Press, Oxford.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Steyer, Kathrin. 2013. *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. (Studien zur Deutschen Sprache 60). Narr, Tübingen.

Steyer, Kathrin and Annelen Brunner. 2014: Contexts, Patterns, Interrelations - New Ways of Presenting Multi-word Expressions. Proceedings of the 10th Workshop on Multiword Expressions (MWE). ACL Anthology. http://www.aclweb.org/anthology/W/W14/W14-0814.pdfshttp://www.aclweb.org/anthology/W/W14/W14-0814.pdf

Internet Sources (Accessed on 7 August 2015)

CII 2015. *COSMAS II. Corpus Search, Management and Analysis System* http://www.ids-mannheim.de/cosmas2/

DeReKo 2015. *Deutsches Referenzkorpus. / Archiv der Korpora geschriebener Gegenwartssprache 2014-II* (Release 11.09.2014). Institut für Deutsche Sprache, Mannheim: http://www.ids-mannheim.de/DeReKo.

SKE: Sketch Engine. https://www.sketchengine.co.uk/

Steyer, Kathrin, Annelen Brunner and Christian Zimmermann. 2013. Wortverbindungsfelder Version 3: Grund. http://wvonline.ids-mannheim.de/wvfelder-v3/

Appendix:

Abbreviations

ADJ: adjective
ADV: adverb
CA: IDS collocation analysis (Belica 1995)
MWE: multi word expression
MWP: multi word pattern
N: noun
UWV: Usuelle Wortverbindungen
V: verb

Figures

Figure 1. Collocation profiles *mit Genugtuung* (cf. CA) and *con satisfacción – with satisfaction* (cf. SKE) (snippet)
Figure 2. Filler tables of *mit – con –* with # (1 slot) *Genugtuung – satisfacción – satisfaction* (snippet)
Figure 3. Trigram filler tables *mit – with – con # # # Genugtuung – satisfaction – satisfacción* (snippet)