A Corpus Based Approach to Near Synonymy of German Multi-Word Expressions*

Christiane Hümmer

Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstr. 22/23,
10117 Berlin, Germany
Email: huemmer@bbaw.de

Abstract. The core of this paper is a detailed corpus-based analysis of the two nearly synonymous German idioms etw. liegt jmdm. im Blut and etw. ist jmdm. in die Wiege gelegt. The central conclusions drawn from this analysis are: On the basis of the behaviour of the semantic arguments of the two idioms – their presence or absence as well as certain semantic properties – clear statements can be made about the context conditions under which the two idioms are interchangeable and those allowing the realisation of one of them while excluding the other one. Furthermore, it is stated that even in the contexts that allow both idioms, the choice of one or the other makes a subtle difference. This difference has to do with the metaphorical image encoded in the idiom. The prominent degree of prototypicality of certain traits demonstrates that speakers actively use these subtle differences. The paper constitutes thus an investigation on the level below WordNet synsets discussing the concept of synonymy underlying WordNet organisation.

1 Introduction

In this paper, the results of a corpus-based analysis of the following two nearly synonymous idioms are reported:

etw. ist jmdm. in die Wiege gelegt	*etw. liegt jmdm. im Blut*
'sth. was put in sb.'s cradle'	'sth. lies in sb.'s blood'

They were taken from a large collection of semantically closely related multi-word expressions (MWE), collected from an onomasiologically sorted [1] and a synonym dictionary of German [2]. The collection of MWEs serves as the data base for a larger (PhD-) project about synonymy of MWEs in German. This paper exemplifies the methodology used and the results that can be expected in this project.

* I would like to thank Patrick Hanks for very patiently helping me work out and edit this paper as well as Kerstin Krell and Ekaterini Stathi for comments on earlier versions. I am also greatly indepted to Christiane Fellbaum for constantly accompanying my work with her advice and for always being open for discussions on any linguistic subject. Work on my PhD dissertation has been made possible by the Wolfgang-Paul-Award of the Alexander-von-Humboldt-foundation, imparted to Christiane Fellbaum. Thanks also to my colleagues Alexander Geyken, Alexej Sokirko und Gerald Neumann, who facilitate the access to the DWDS corpus and provide the computational tools for corpus search and to Ralf Wolz for editing advice.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 142–149.

The PhD dissertation is part of the project "Kollokationen im Wörterbuch" ('Collocations in the Dictionary') at the Berlin-Brandenburg Academy of Sciences. This project aims at investigating syntactic, semantic and morphological properties of German idioms on the basis of the 980 M word DWDS corpus, which was compiled from texts representing a wide variety of genres and covering the entire 20th century [3]. The same corpus was used as a source of empirical evidence for the investigation presented here. For each of the MWEs examined, a subcorpus containing all the occurrences of that expression in the corpus has been extracted using the linguistic query tools developed by members of the project "Kollokationen im Wörterbuch". All the conclusions drawn and all the quantitative statements made in this paper are based on a manual analysis of the whole subcorpora.

Since idioms are not generally found in WordNet and constitute a number of problems for codification (see Fellbaum 1998 [4] and Fellbaum 2002 [5]), the two idioms examined are not WordNet entries. Nevertheless, they are semantically close enough to be assumed to be candidates for membership in one common synset once they can be enconded. The lexicological case study presented in this paper is therefore a suitable example for investigating the concept of synonymy WordNet makes use of. As Miller et al. [6] point out, the WordNet organisation principle of synonymy is based on the idea of substitutability without change of truth values. Since the authors doubt the existence of absolute synonyms, substitutability in some contexts is assumed as a sufficient prerequisite for making two lexical units become members of the same synset. This paper shows the results of basing this intuition on a corpus-based research.

2 Results

Conclusions that have been drawn from the corpus data focus on two main questions:

1. What are the context conditions that make the two idioms converge or diverge semantically?
2. Assuming that even interchangeable expressions are not absolutely synonymous, what governs the choice between the two idioms in contexts where they are interchangeable? And how can this be identified in the corpus data?

Concerning the first question, it can be said that the conditions of semantic convergence and divergence can be formulated quite clearly in terms of the behaviour of the semantic arguments of the idioms.

As for the external valency of the idiom *etw. ist jmdm. in die Wiege gelegt*, its maximal realisation is achieved when the idiom is realised in the active voice. Although this is not its prototypical[1] syntactic form, it occurs in a considerable amount of cases (150 out of 609, see below), as corpus evidence proves. This maximal realisation can be classified as an instance of the semantic frame

DONOR – RECIPIENT – THEME[2],

[1] The notion of prototypicality of meaning and form, very important for the present investigation, is also found in many previous publications, e.g. Hanks 1994 [7] and Hanks 1997 [8].

[2] The names of the semantic roles are taken from the FrameNet specifications of Frame Elements ('Giving Frame') [9]

syntactically realised as subject – indirect object – direct object.
The idiom *etw. liegt jmdm. im Blut* takes arguments that may be described as

PROPERTY and PROTAGONIST[3],

syntactically realised as subject and possessive dative. A SOURCE is also often mentioned but not as part of the argument structure of the idiom.

In order to gain an overview of corpus evidence, for each idiom a table was constructed with columns as slots for the arguments and the rows containing the particular lexical items filling the semantic argument positions in the corpus (see Table 1 for a very small part of this table).

Table 1. Extract from the table containig realisations of semantic arguments of *etw. liegt jmdm. im Blut*

	PROTAGONIST	PROPERTY	SOURCE
1.	Jason Gebert (poss. dat.)	eine Freude an schönen Farben … (Subj) (‘taking pleasure in beautiful colours…’)	vom Vater her (Adjunct/VP) (‘from his father’)
2.	ihr (poss. dat.) (‘to her’)	Die Fliegerei (Subj) (‘flying’)	Vater war Flugkapitän … (context) (‘father was an aircraft captain’)
3.	denen (poss. dat.) (‘to them’)	das Bedürfnis nach Bewegungsfreiheit… (Subj) (‘a want for freedom of movement’)	als Britensprösslingen (Adjunct/Dat) (‘as offsprings of British’)
4.	den Clintons (poss. dat.) (‘to the Clintons’)	Wahlkämpfe (Subj) (‘election campaigns’)	[den Clintons (Dat]
5.	den Katholiken (poss. dat.) (‘to the catholic’)	das Lügen (Subj) (‘lying’)	[den Katholiken (Dat)]
6.	euch Bienen (poss. dat.) (‘to you bees’)	Fliegen und immer fliegen (Subj) (‘flying and always flying’)	[euch Bienen (Dat)]
7.	ihm (poss. dat.) (‘to him’)	Das unsittliche Leben (Subj) (‘the immoral life’)	
8.	[deutschen (Adj modifying Blut)] (‘German’)	Die Angriffslust (Subj) (‘aggressiveness’)	deutschen (Adj modifying Blut)

These tables show that basic semantic frames are frequently modified considerably in actual use:

In the context of the expression *jmdm. liegt etw. im Blut*, very frequently expressions can be found that encode the SOURCE of the PROPERTY attributed to the PROTAGONIST. This SOURCE is, in most cases, the family or a group (often genetically specified) that the

[3] FrameNet roles from the ‘Mental_property’ frame [9]

PROTAGONIST belongs to. It appears as merged with the entity denoting the PROTAGONIST (rows 4,5,6,8), or as an independent expression in the closer (rows 1,3) or wider context (row 2).

This SOURCE argument is very similar to the DONOR semantic role of *etw. ist jmdm. in die Wiege gelegt*. The difference between the two lies in the emphasis that can be given to this argument: the DONOR in *etw. ist jmdm. in die Wiege gelegt* can be expressed as an Agent taking the subject position, whereas SOURCE can only appear in less prominent positions.[4] In addition, the fact that SOURCE is often explicitly expressed in the context of *etw. liegt jmdm. im Blut* assigns a role to the PROTAGONIST to whom a PROPERTY is attributed, similar to the RECIPIENT role of the idiom *etw. ist jmdm. in die Wiege gelegt*.

The expression *etw. ist jmdm. in die Wiege gelegt* is realised in 460 cases out of 609 in a combination of passive voice and past tense or in a special German passive form called 'Zustandspassiv'. As a result, in many cases DONOR is not expressed in the context. Instead of an activity, the predicate denotes then a state in which THEME is a PROPERTY of the RECIPIENT. In such cases THEME and PROPERTY are very similar to RECIPIENT and PROTAGONIST in *etw. liegt jmdm. im Blut*.

> from: Frankfurter Rundschau 09.03.2000, S. 12
> Hilfsbereitschaft und der Blick für Missstände und Ungerechtigkeit scheinen der tatkräftigen Frau *in die Wiege gelegt.*
> ('helpfulness and an eye for problems and injustice seem to *have been put in the cradle* of this energetic woman')

The elements that normally fill the argument positions for THEME in *etw. ist jmdm. in die Wiege gelegt* and for PROPERTY in *etw. liegt jmdm. im Blut* to a large extent stem from the same semantic class. As can be seen in the following examples, they are often very similar to each other:

Table 2. Examples of similar lexical items filling the Property argument of *etw. liegt jmdm. im Blut* and the Theme argument of *etw. ist jmdm. in die Wiege gelegt*

PROPERTY argument of *etw. liegt jmdm. im Blut*	THEME argument of *etw. ist jmdm. in die Wiege gelegt*
Das Verkaufen (Faculty) ('selling')	Millionen von Kuscheltieren in alle Welt zu verkaufen (Faculty) ('to sell millions of cuddly toys to the whole world')
die Liebe zu alter Technik (Inclination) ('a love for old technology')	die Liebe zur Musik (Inclination) ('a love for music')
Opposition (Attitude) ('opposition')	Widerstand (Attitude) ('resistance')

[4] With Grimshaw [10] it is assumed that the syntactic function the arguments of a predicate fulfill is determined by an hierarchy of thematic roles and a salience hierarchy of aspectual prominence of arguments. In particular, Grimshaw assumes that the thematically and aspectually most prominent argument is always realised as the subject.

From these observations, general contextual conditions can be formulated under which the two idioms converge semantically.

Etw. ist jmdm. in die Wiege gelegt converges maximally with *etw. liegt jmdm. im Blut* when it is realised in the 'Zustandspassiv', therefore leaving out the DONOR semantic argument and expressing a state instead of an activity. In addition, the THEME argument is filled by a lexical item that can be categorised as belonging to the semantic class of Faculties, Inclinations and Attitudes.

Etw. liegt jmdm. im Blut converges with *etw. ist jmdm. in die Wiege gelegt* when it takes an additional semantic argument similar to the DONOR of *etw. ist jmdm. in die Wiege gelegt.*

In such cases, one expressions can be substituted for the other salva veritate.

The opposite case, where the context only allows one of the idioms, can be described as the complement of what was said above. Basically, *etw. liegt jmdm. im Blut* cannot be substituted for *etw. ist jmdm. in die Wiege gelegt* when *etw. ist jmdm. in die Wiege gelegt* is realised in the active voice as an activity of some Agent that takes the semantic role of a DONOR. Another condition that makes the two idioms diverge is fulfilled when the filler for the THEME position belongs to a semantic class that is not compatible with the PROPERTY position and vice versa. For example, something that has been put in somebody's cradle has to be interpreted as a condition capable of affecting the whole life of that person from the beginning on:

> from: Frankfurter Rundschau 20.10.1997, S. 18
> Der Sohn eines Gummisohlenfabrikanten, dem *weder Geld noch Kunstwerke in die Wiege gelegt worden waren*, hatte sich als junger Mann in österreichischen Revuen als Werber für Schuhcreme verdingt . . .
> ('The son of a maker of rubber soles, *into whose cradle neither money nor works of art had been put*, hired himself out, as a young man, to Austrian revues as an advertiser for shoe polish. . . ')

In this context, it is impossible to use the idiom *etw. liegt jmdm. im Blut*.

In other words, presence or absence of certain semantic arguments as well as the semantic and syntactic role they play in the sentence and the semantic class of the lexical items that realise them determines closeness or distance of the two nearly synonymous expressions.

Concerning the question what makes the two idioms different in contexts where they are maximally synonymous, the focus of the investigation was placed on the influence of the metaphoric images and connotations associated with them.

As a starting point, it can be said that *blood* is a much stronger and more drastic image than *cradle*. When *etw. liegt jmdm. im Blut* is used, the speaker usually makes either a very strong statement about the deep-rootedness of a PROPERTY in a PROTAGONIST or it serves to express his (ironic) distancing himself from what he says. This happens above all in cases where a cliché is expressed, which is very frequently the case with *etw. liegt jmdm. im Blut* (see below):

> From: Frankfurter Rundschau (Jahresausgabe 1998)
> Wenn sie nur nicht so aggressiv wären, die Rothäute mit ihren Hakennasen und der Kriegsbemalung. Dabei steht ihnen doch der Federschmuck, den sie stets tragen, so gut. Und wenn sie erst ihre berühmten Tänze aufführen. Großartig. Wir alle wissen

> doch: *Negern und Indianern liegt der Rhythmus im Blut*. Das ist einfach angeboren bei den schwarzen Perlen und roten Kriegern. Wie – Sie finden diese Sätze rassistisch, dumm und unerträglich? Wir auch! Wir fragen uns nur, warum immer öfter als Werbe-Gag vielerorts mannshohe Abbilder jener Menschen herumstehen, die die doch so zivilisierten Weißen vor noch nicht langer Zeit versklavt, verfolgt und ermordet haben.
> ('if only they wouldn't be so aggressive, those redskins with their hooknoses and their war paint. And the feathered headdress they wear fits them so well. And when they perform their famous dances. Grandiose. We all know of course: Rhythm is in the blood of the Negroes and the Indians. It's simply innate in those black pearls and red warriors. What? You find these sentences racist, stupid, and unbearable? So do we! We just ask ourselves why more and more often, as a commercial gimmick, in some places a life-sized picture of those people, who have been enslaved, persecuted and murdered by those remarkably civilised whites not long ago, stands around.')

This assumption is strongly supported by some observations from the corpus.

For example, typical modifications taken by the idioms can give a hint of their characteristic semantic traits. Under this view, the fact that *tief* ('deeply') is a typical modification that appears with the idiom *etw. liegt jmdm. im Blut* (8/326; MI: ∼2.74) highlights the profound rootedness of the PROPERTY in the PROTAGONIST expressed by the idiom. In contrast to this, a typical modification of *etw. ist jmdm. in die Wiege gelegt* is *bereits/schon* ('already') (80/609; MI: ∼1.9), emphasising the early age of the RECIPIENT when receiving THEME.

Another fact related to the meaning of the image is the prototypicality of having a genetic group or an individual representing a genetic group in the PROTAGONIST position of *etw. liegt jmdm. im Blut*. This happens in 78 out of 326 cases, not counting those cases where the PROTAGONIST position is filled by a pronoun whose reference cannot be recovered from the context of one sentence. Some examples are: *die Deutschen* ('Germans') (10 times), *die Schweizer* ('Swiss') (2 times), *Neger* ('Negroes') (2 times) *Indianer* ('Indians'), *Latinos* (2 times), *Juden* ('Jews') (3 times), *Briten* ('British') (2 times) *Südländer* ('southerners') etc.

Altogether, prototypicality effects seem to support the intuitive insight that the image carried by *etw. liegt jmdm. im Blut* favours the interpretation of the PROPERTY argument as something innate while the THEME argument of *etw. ist jmdm. in die Wiege gelegt* tends to be interpreted as something determined by social circumstances or education.

3 Conclusion

The discussion of the corpus-based analysis of two nearly synonymous idioms in this paper shows three main points:

Semantic convergence and divergence of the two idioms is proportional to the behaviour of their semantic arguments. The idiom *etw. ist jmdm. in die Wiege gelegt* was basically analysed as belonging to the frame DONOR (Subject) – RECIPIENT (IO) – THEME (DO), but it converges in its semantic interpretation with *etw. liegt jmdm. im Blut* under the following conditions:

- the DONOR is not present (basically when the idiom is realised in the passive);

- the 'Zustandspassiv' changes the interpretation of the idiom: It encodes a state instead of an activity and the THEME arguments can be seen as a PROPERTY of a PROTAGONIST (otherwise known as the RECIPIENT);
- the lexical material filling the THEME argument position can be interpreted as belonging to the semantic class of faculties, inclinations or attitudes.

It diverges most strongly from *etw. liegt jmdm. im Blut* when

- the subject is present and interpreted as an agent that carries out a giving action or when
- the lexical material in the THEME position is to be interpreted as an starting condition for some individual (the PROTAGONIST) from the beginning of his life.

The idiom *etw. liegt jmdm. im Blut* was basically analysed as belonging to the frame

PROPERTY (subject) – PROTAGONIST (IO).

It converges with and diverges from *etw. ist jmdm. in die Wiege gelegt* mainly with the presence or absence of an additional semantic argument. This argument contains information about the source (or DONOR) of the PROPERTY.

Even in contexts where both idioms should be equally possible the choice of one or the other makes a subtle difference that has to do with the idiomatic image associated with the idiom.

Blut ('blood') is a much stronger image than *Wiege* ('cradle'). In consequence, the use of *etw. ist jmdm. in die Wiege gelegt* is, in most cases, more neutral than *etw. liegt jmdm. im Blut*. When *etw. liegt jmdm. im Blut* is used, it frequently implies either a much more radical statement about the deep-rootedness of a PROPERTY in some PROTAGONIST or, to the other extreme, serves as a way of marking an ironic distancing of the speaker.

The fact that in the corpus the PROTAGONIST can denote a genetic group in both *etw. ist jmdm. in die Wiege gelegt* and *etw. liegt jmdm. im Blut*, but is realised as such very significantly more frequently with *etw. liegt jmdm. im Blut* is only one sign that shows how speakers make use of this distinction.

In summary, from the fact that passive and 'Zustandspassiv' are prototypical syntactic forms for *etw. ist jmdm. in die Wiege gelegt* and that a genetic group in the PROTAGONIST position is prototypical for *etw. liegt jmdm. im Blut* the conclusion can be drawn that the two idioms converge significantly in the language use.

Still, corpus evidence demonstrates that speakers agree on a subtle semantic difference between the two.

With respect to synonymy, the case study at hand supports the claim that the intuitive notion of substitutability should be grounded on corpus evidence. Generalisations over corpus data allow insight on the degree of synonymy in terms of shared or mutually exclusive context conditions as well as about preferred or prototypical contexts for the realisation of the two synonym candidates. Such statements are very important for tasks such as fine-grained lexical choice for Natural Language Generation. Parallel to what Edmonds [11] and Edmonds and Hirst [12] show for synonym words on the basis of dictionary definitions, it can be said that for those tasks a much more fined-grained distinction between lexical units is needed than the one provided by WordNet synsets.

References

1. Hessky, R., Ettinger, S. (eds.): Deutsche Redewendungen. Ein Wörter- und Übungsbuch für Fortgeschrittene. Narr, Tübingen (1997).
2. Schemann, H. (ed.): Synonymenwörterbuch der deutschen Redensarten. Unter Mitarbeit von von Renate Birkenhauer. Straelener manuskripte, Straelen (1989).
3. Cavar, D., Geyken, A., Neumann, G. Digital Dictionary of the 20th Century German Language. In: Erjavec, T., Gros, J. (eds.): Proceedings of the Language Technologies Conference 17–18 october 2000 Ljubljana. On-line proceedings (2000) `http://nl.ijs.si/isjt00/index-en.html`.
4. Fellbaum, C.: Towards a Representation of Idioms in WordNet. In: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems. University of Montréal, Montréal, Canada (1998) 52–57.
5. Fellbaum, C.: VP Idioms in the Lexicon: Topics for Research using a Very Large Corpus. In: Busemann, S. (ed.): Proceedings of KONVENS 2002 30. september–2. october 2002 Saarbrücken. On-line Proceedings (2002) `http://konvens2002.dfki.de/cd/index.html`.
6. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J.: Introduction to WordNet: An On-Line Lexical Database. Journal of Lexicography 3(4) (1990) 235–244.
7. Hanks, P.: Linguistic Norms and Pragmatic Exploitations. Or, why Lexicographers need Prototype Theory, and Vice Versa. In: Kiefer, F., Kiss, T., Pajzs, J. (eds.): Papers in Computational Lexicography: Complex 1994. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest (1994) 89–113.
8. Hanks, P.: Lexical Sets: Relevance and Probability. In: Lewandowska-Tomaszczyk, B., Thelen, M. (eds.): Translation and Meaning. Part 4. Euroterm, Maastricht (1997).
9. FrameNet: `http://www.icsi.berkeley.edu/~framenet/`.
10. Grimshaw, J.: Argument Structure. MIT Press, Cambridge, Mass London, England (1990).
11. Edmonds, P.: Semantic Representation of Near-Synonyms for Automatic Lexical Choice. University of Toronto, Toronto (1999).
12. Edmonds, P., Hirst, G.: Near-Synonymy and Lexical Choice. Computational Linguistics 28(2) (2002) 105–144.