Dynamic lexicographic data modelling
A diachronic dictionary development report

Paul Gévaudan and Dirk Wiebel

SFB 441 Linguistische Datenstrukturen
Nauklerstr. 35, 72074 Tuebingen, Germany
{paul.gevaudan, dirk.wiebel}@uni-tuebingen.de

Abstract

Lexical units of several different language areas show noticeable similarities in their semantic structure, corresponding to their etymological development. This phenomenon of *polygenetical evolution* leads to the assumption that lexical innovations are strongly influenced by cognitive constants. These cognitive constants can be seen as a result of anthropological predispositions. In order to examine the relevant constants and the resulting polygenesis, a model of diachronic *filiation* has been developed. This model has the capacity to analyse highly complex cases of lexical evolution. An easily readable access and representation system is given by a linear notation method. This method is able to annotate all lexical innovations; however, it is not suitable for computer based analyses. We have therefore subsequently introduced an entity-relationship model representing the linguistic data model. Integrated into a powerful DBMS, this allows a dynamic representation of the underlying diachronic lexicon which has been compiled in our projects.

Polygenesis and diachronic data

1 Subject and goals

The main subject of our research projects is the diachronical development of lexical onomasiological data, reduced to the domain of body part designations. One of the goals is to explain, if the development of these expressions is based on polygenetical paths and if these paths are mainly influenced anthropologically or even genetically. Currently, there are two different projects dealing with an equivalent data structure. DECOLAR (Dictionnaire Etymologique et Cognitif des Langues Romanes) is working on the domain on body part designations in general, reduced to Romance languages. $\text{LexiType}_{\text{Dia}}$ (= Project B6 in our Collaborative Research Centre 'Linguistic Data Structures') deals with a larger sample of about 50 languages[1] located in widely different areas, but with a reduced subset of parts of the human head. Crucially for the topic of polygenesis this large variety makes it possible to exclude language contact phenomena.

1.1 Polygenesis

Approaches based on typology or on theories of nativeness usually have their fundaments in polygenetical facts about language change. Polygenetical processes are diachronical processes which are similar across several different languages or language stages. Interactive influence must be excluded for this phenomenon. Therefore, the problem of pointing out polygenesis is a typical data problem. Methodological and theoretical reflections as well as empirical data analysis form the basis of our research. Especially the main question of whether polygenesis occurs due to anthropological or genetical predisposition has to be answered empirically. For the empirical analysis, the choice of the model is of fundamental importance for the final analysis.

[1] Among others: Albanian, Chinese, English, German, Hungarian, Japanese, Maori, Nahuatl, Quechua, Russian, Scottish Gaelic, Swahili, Tamil, Vietnamese, Warlpiri, Zeltal, etc.

1.2 Methods and theory

The analysis is based on a lexical and diachronical model of 'filiation', which distinguishes different types of lexical genesis (changes of meaning, changes of designation, word morphology or loan words). In order to allow statistical methods, these distinctions have to be standardised.

One of the most important distinctions has to be made between changes of meaning (semasiological perspective) and changes of designation (onomasiological perspective). The onomasiological perspective always allows a general starting point for every type of change in lexical or grammatical meaning, because every innovation in a vocabulary immediately involves a new designation for a new concept. Therefore, both projects offer an onomasiological access to the system. The lexical material is acquired on a basis of concepts; subsequent steps include the genesis of the denotation.

1.3 Data acquisition and maintenance

The lexical material which forms the empirical basis for our research consists mainly of secondary data, which was acquired from diachronical and synchronical dictionaries, glossaries and language descriptions. These data do not provide natural language directly; nevertheless they are necessary to offer comparability for

- different languages within different language families,
- different types of genesis, and
- diachronic and synchronic data.

Besides the labour-intensive acquisition of multi-areal language data, the large size of the data and its complexity make great demands on the underlying structure and the management system. This structure has to meet several requirements. On the one hand, it has to offer sufficient possibilities of distinguishing data regarding their genesis within largely different languages or language families as well as the descriptions in synchrony and diachrony. On the other hand, a comparable data model has to be created, in which the data can be easily opposed to approximate items. Apart from these requirements, the whole system

has to be made accessible in a human-readable form in order to identify and to prove the items, their structures and their relations. It is therefore necessary to create a minimal but sufficient feature-based descriptional system, which will be described briefly in the following paragraphs.

2 Linguistic Background

2.1 Data selection and semiotic perspective shifting

The methodology of our data selection and analysis comprises four steps which can be described by means of semiotic perspectives. This procedure involving the four possible semiotic perspectives is shown in the following figure:

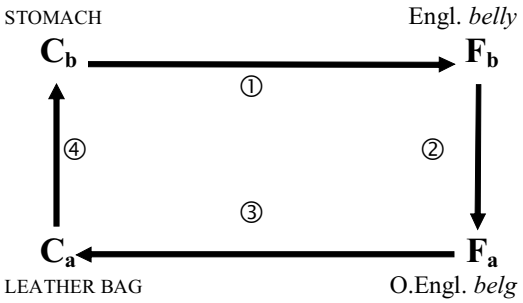

Fig. 1: The four-step-analysis-method of the LexiType$_{Dia}$ and the DECOLAR project

① Onomasiological 'start': In the explored synchrony, a given concept C_b (e.g. STOMACH) is expressed by a certain form F_b (e.g. Engl. *belly*). As mentioned above the onomasiological selection assures our language sample to be comparative.
② Diachronical retrospection: Since F_b (*belly*) has become a new expression for the concept C_b at a certain moment, we have to examine in a retrospective-diachronical semiotic perspective from which lexical antecedent it originates from. In our case, the antecedent (F_a) of the form Engl. *belly* is O.Engl. *belg*, from which F_b is derived by suffixation.
③ Semasiological description of an antecedent synchrony: In this step, a semasiological perspective can be established. The form F_a expresses the concept C_a (i.e. LEATHER BAG).
④ Conceptual analysis: Here we identify the cognitive semantic relations between C_a and C_b (here: 'metaphoric similarity').[2]

2.2 Three-dimensional cross-classification based on the filiation model of lexical evolution.

The diachronic analysis of designations is based on the *filiation model* (Gévaudan 2002, 2003, ms.), which takes into consideration all kinds of lexical evolutions and offers standardised explanations of phenomena like semantic change, word formation and borrowing, which represent the three possible ways of enriching the vocabulary of a language. Given that they are not only of lexicologic or diachronic interest, semantic change, word formation and borrowing have so far been treated in different linguistic disciplines, producing incompatible results with regard to the analysis of lexical evolution.

To illustrate the problem, consider the examples below, which represent three different verbalisation strategies for the designation of the concept SWEET PEPPER (*capiscum frutescens*) in Spanish, French, and Hungarian:

(1) Sp. *pimiento* ''sweet pepper' ←Sp. *pimiento* 'pepper'
(2) Fr. *poivron* 'sweet pepper' ←Fr. *poivre* 'pepper'
(3) Hung. *paprika* 'sweet pepper'←Serb-Cr. *pàpar* 'pepper'

It is eye-catching that the semantic processes underlying these innovations are identical. Nevertheless, neither the theories of word formation nor those dealing with borrowing would describe this semantic innovation as theories of semantic change would do. In addition to that, the approaches dealing with semantic change describe these processes by means of holistic categories which delimitate other types of lexical innovation which involve evidently comparable semantic processes.

The filiation theory, however, provides a standardised explanation and method of analysis for all the different kinds of *lexical innovation*. The analysis is based on a three-dimensional, cross-classificational grid which involves a semantic, a morphological and a stratificational dimension, as it is shown in fig. 2.

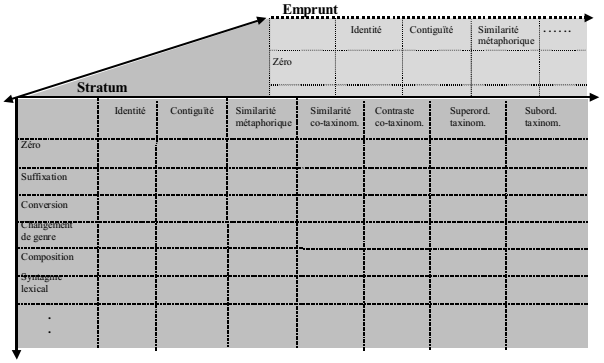

Fig. 2: Three-dimensional cross-classificational grid of filiation

The semantic level of the analysis provides categories like 'identity', 'contiguity' (the motivation of metonymy), 'metaphoric similarity' (the motivation of metaphor), etc. (cf. note 1). Contrary to traditional semantic approaches which deal only with morphological continuity, i.e. unchanged forms, these categories do not presume anything at the level of morphology, where we are dealing with categories like 'zero' (unchanged forms), 'change of number', 'conversion', 'suffixation', 'prefixation', 'compound', 'lexical phrase' (collocation), ellipsis, etc.

[2] The set of cognitive-semantic relations used within the LexiTypeDia and the DECOLAR project comprises: 'identity', 'contiguity' (the motivation of metonymy), 'metaphoric similarity' (the motivation of metaphor), 'generalisation' / 'taxonomic superordination' (the motivation of semantic extension), 'specialisation' / 'taxonomic subordination' (the motivation of semantic narrowing), 'co-taxonomic similarity' (the motivation of co-hyponymic transfer). For a definition of these categories by means of cognitive linguistics, cf. Blank (1997, 2003) Koch (1999, 2001a, 2001b), Gévaudan (2002, 2003, ms.).

2.3 Readability and linear notation

Crossing the semantic and morphological level yields to bi-parametric results as it is the case in the following example:

(4) Alb. *kokërdhok* 'eyeball'
<metaphoric Similarity.Suffixation<
Alb. *kokërr* 'knob'

Here, the result of the cross-classification is expressed in a linear expression with the following pattern:

(5) successor
<[semantic filiation].[morphological filiation]<
antecedent

The linear notation allows us to show more complex cross-classificational constellations in a clear way. This already shows the three-dimensional analysis, whose graphical representation is not easy to provide. The three-dimensional analysis of the above examples (1)–(3) points out the advantages of the applied linear notation:

(1') Sp. *pimiento* 'sweet pepper'
<co-taxonomic similarity.zero.stratum<
Sp. *pimiento* 'pepper'
(2') Fr. *poivron* 'sweet pepper'
<co-taxonomic similarity.suffixation.stratum<
Fr. *poivre* 'pepper'
(3') Hung. *paprika* 'sweet pepper'
<co-taxonomic similarity.suffixation.loaning<
Serb-Cr. *pàpar* 'pepper'

2.4 Multiple and paradigmatic filiation: some more complex cases

The linear notation also allows to represent what we call *multiple filiation* (cf. Gévaudan 1999, ms), i.e. the multi-factor-analysis of compounds and lexicalised phrases. Consider the following example:

(6) Est. *pöseluu* 'cheekbone'
← Est. *luu* 'bone' + Est. *pösk* 'cheek'

In this case, we have to take into account that the resulting lexical unit has not only one, but two antecedents, which are both in a certain semantic relation to the whole construction. This is shown in the following figure:

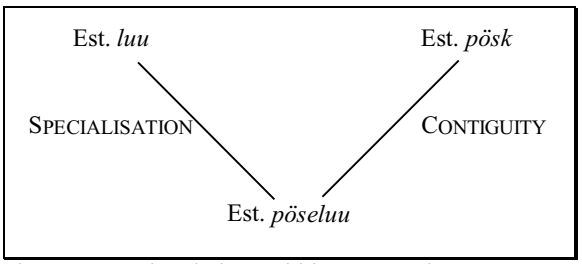

Fig. 3: Semantic relations within est. *pöseluu*

While at the morphological and stratificational level we still have a holistic classification of the lexical innovation, we must describe a multiple filiation at the semantic level. Again, the simplest way to represent the multiple filiation is the linear notation:

(6') Est. *pöseluu* 'cheekbone'
<specialisation+contiguity.compound<
Est. *luu* 'bone' + Est. *pösk* 'cheek'

As described in Gévaudan (1999), the semantic filiation of lexicalised phrases (collocations) is perfectly comparable with that of lexical compounds. This is shown in the following example:

(7a) Russ. *glaznoe jabloko* 'eyeball'
<contiguity+metaphSimilarity.compound.stratum<
Russ. *jabloko* 'apple' + Russ. *glaz* 'eye'
(7b) Germ. *Augapfel* 'eyeball'
<contiguity+metaphSimilarity.compound.stratum<
Germ. *Apfel* 'apple' + Germ. *Auge* 'eye'

While lexicalised phrases have the morphological structure of syntactic phrases, lexical compounds have a word formation specific morphology. Preferences on forming and/or lexicalising phrases or compounds seem to depend on structural predispositions. This is one of the arguments sustaining Koch's (2001b) claim for a *lexical typology*.

On top of the multiple filiation, another kind of complex filiation, which we described as *paradigmatic filiation* (cf. Gévaudan ms.), represents the phenomenon of *calque*, usually known as 'loan translations' and/or 'semantic loans'. It consists in an imitation of the semantic structure of a foreign word, or of a lexical paradigm. The lexical unit Fr. *bassin* 'pelvis' e.g. is based upon Fr. *bassin* 'basin' – this lexical innovation imitates the Latin constellation where the anatomic designation *pelvis* has been metaphorically derived from Lt. *pelvis* 'basin'. In linear notation, this case is described as follows:

(8) Fr. *bassin* 'pelvis'
<metaphSimilarity.zero.**calque**<
Fr. *bassin* 'basin'
:: Lt. *pelvis* 'pelvis'← Lt. *pelvis* 'basin'

Both phenomena, multiple and paradigmatic filiation, can occur within the same case of innovation:

(9) Germ. *Wirbelsäule* 'vertebral column'
<metaphSimilarity+contiguity.compound.calque<
Germ. *Säule* 'column' + Germ. *Wirbel* 'vertebra'
:: Lt. *columna vertebralis* 'vertebral column'
← Lt. *columna* 'column' + Lt. *vertebra* 'vertebra'

At this place, we will not discuss further cases of complex filiation, e.g. phenomena like popular etymology, analogical change or antonomasia.

3 Data Design

The linear notation offers an easily readable user interface, which can describe the whole complexity of the cross-classificational grid. However, it does not offer an efficient possibility for data storage, maintenance or access.

The model of lexical filiation provides a formal representation and a standardised classificational system for all possible cases of lexical evolution. Its analysis is

based on the three-dimensional cross-classificational grid (see fig. 2) which allows a consistent combination of semantic, morphological and stratificational criteria. For instance, to enable a cross-linguistic search based on the filiation sequence on the level of concepts, the design of the underlying data structure has to comply with the whole complexity of our lexicographical filiation model.

On the technical side, a fast and reliable storage system suitable for multi-user access had to be to be equipped with an expandable collection of user interfaces and maintenance tools. The support of several export formats based on one central data source was required in order to enable usability for different types of publication media. The data and a large set of tools have been integrated into the TUSNELDA collection (Tuebinger Sammlung Nutzbarer Empirischer Linguistischer Datenstrukturen, 'Tuebingen collection of usable empirical linguistic data structures') of our Collaborative Research Centre (Wagner & Kallmeyer 2001). Its main purpose is to solve the central problem of transferring synchronic data units and their relations into a diachronic representation.

Our data are structured as synchronic entities and diachronic relations. This lent itself to a relational data model, based on the standard entity-relationship scheme. The huge complexity of the lexical diachronic configurations explains the need for a powerful and extensible database management system (DBMS). The large variety of the language samples in project B6 and the need for recursive queries were another argument for a professional DBMS. In order to ensure data integrity while using a very complex model, we have chosen to integrate most of the query functions into the level of the DBMS. Because of these requirements, the initial data collection was migrated to IBM's DB2 system during the modelling period.

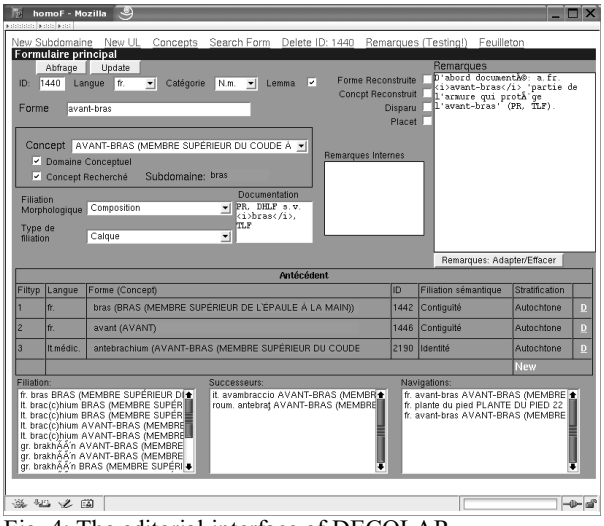

Fig. 4: The editorial interface of DECOLAR

The first experiences during the period of data collection showed that proprietary DBMS applications are insufficiently expansible e.g. for our need for recursive queries with a markup of diachronic levels and morphological types of antecedents. Therefore, our approach was to implement the system with standardised query, program, and markup languages, and to build the interfaces using mainly open source tools. For data storage and maintenance, SQL is the standardised language within the DBMS. The user interfaces are written in PHP, which offers powerful DBMS functions and which can be easily integrated into HTML as a markup language. Furthermore, HTML offers the advantage of being standardised and therefore platform-independent. PHP gives the possibility of exporting data into PDF or XML. Interfaces for XML based output will be developed in the future. A preliminary version of a guest user interface is available at http://www.sfb441.uni-tuebingen.de/b6/. The final version of the online dictionaries DECOLAR and B6 ($LexiType_{Dia}$) will be published in summer 2004.

References

Blank, A. (1997). Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen. Tübingen: Niemeyer.

Blank, A. (2003). Words and concepts in time: Towards cognitive onomasiology. In Chr. Schwarze / R. Eckardt / K. Heusinger (eds.), Words in Time. Diachronic Semantics from Different Points of View, Berlin; New York.

Gévaudan, P. (1999). Semantische Relationen in nominalen und adjektivischen Kompositionen und Syntagmen. In PhiN.Philologie im Netz 9, 11–34. [http://www.fu-berlin.de/phin]

Gévaudan, P. (2002). Fondements sémiologiques du modèle de la filiation lexicale. In PhiN.Philologie im Netz 22, 1–26. [http://www.phin.de/phin/]

Gévaudan, P. (2003). Lexikalische Filiation. Eine Synthese von Onomasiologie, Semasiologie und Etymologie. In Blank, Andreas / Koch, Peter (ed.), Kognitive romanische Onomasiologie und Semasiologie (pp. 189–211). Tübingen.

Gévaudan, P. (ms). Klassifikation lexikalischer Entwicklungen. Semantische, morphologische und stratische Filiation. PhD-Thesis, University of Tübingen.

Koch, P. (1999). Frame and contiguity: On the cognitive basis of metonymy and certain types of word formation. In Panther/Radden (ed.), Metonymy in Language and Thought (pp. 139–167). Amsterdam/Philadelphia.

Koch, P. (2001a). Metonymy: Unity in diversity. Journal of Historical Pragmatics 2/2, 201–244.

Koch, P. (2001b). Lexical typology from a cognitive and lingustic point of view. In M. Haspelmath et al. (eds.), Typology and Language Universals. Berlin/New York: de Gruyter.

Wagner, Andreas and Laura Kallmeyer (2001). Der TUSNELDA-Standard: Ein Korpusannotierungsstandard zur Unterstützung linguistischer Forschung. In Proceedings of GLDV-Frühjahrstagung, Gießen, März 2001 (pp. 253–262). Gießen.