Coding semantic categories cross-linguistically – a challenge for typological research

Christiane von Stutterheim*

Heidelberg University

Article Info

Article history:
available online: July 23, 2021

Article language:
English

keywords:
cross-linguistic research
motion event encoding
space and time

The goal of the paper is to provide a guide for dealing with challenges which occur when dealing with multilingual data at the semantic level. Three dilemmas are pointed out which relate a) to decisions on the size of coding units, b) to the integration of parts of meaning deduced from inference and c) to the diversity of theoretical categories used for describing semantic concepts. Problems and possible solutions are illustrated by looking at two semantic domains: space and time. Recent cross-linguistic studies on spatial and temporal categories are taken as examples for suggesting a four step model for multilingual coding.

1 Introduction

The goal of this paper is to chart some of the problems encountered in data-based cross-linguistic studies and discuss solutions, as well drawbacks, when coding multilingual data at the level of semantics. Focus is placed on language-specific variation across the semantic concepts underlying linguistic expressions.**

An investigation of this issue only makes sense to the extent to which the semantic concepts underlying the different expressions studied are kept constant across languages. As always in empirical studies, it may be necessary to make some compromises in specific cases, but it would obviously not make sense to examine how, for example, "progressive aspect", is expressed in languages A, B and C if this term means something different in all three languages. Although *Pierre est en train de faire le déjeuner* 'Pierre is in the process of preparing lunch' as well as *Peter is preparing lunch* both express some form of progressivity, it would be misleading to code both of them under the same label because in actual fact they express something quite different given marked cross-linguistic differences, as closer examination reveals. Fine-grained differences may be irrelevant for some purposes but they are crucial in achieving a deeper analysis of cross-linguistic variation.

So how can we ensure that the underlying semantic category is kept as constant as possible and relevant for the question at hand? Essentially, there are two ways to proceed: we can characterize the category in a language-dependent or in a language-independent mode. A typical example for a language-dependent mode of analysis is given with definitions such as "progressive aspect indicates that the action is ongoing" or "has not yet come to an end." In fact, such definitions are very often used and are useful as an initial point of departure. But as the example above shows they are often too coarse, and are often biased towards a particular language (typically English).

A less language-dependent process of analysis is generally preferable and can be achieved in various ways. For example, it could involve a neutral definition such as "the concept 'pastness' expresses the relation $t_{sit} < t_0$, where t_{sit} is the time of the situation described by the sentence and t_0 is the time at which the sentence is uttered". Relational definitions of this kind, as often used in formal semantics, are very precise. Although they are helpful in many cases for cross-linguistic purposes they also raise problems: First of all they are still language-dependent, in par-

* Correspondence address: C.v.Stutterheim, Institut für Deutsch als Fremdsprachenphilologie, Plöck 55, D 69117 Heidelberg. e-mail: stutterheim@idf.uni-heidelberg.de

** I am grateful to Mary Carroll, Monique Lambert, Michael Herweg and Wolfgang Klein for substantial comments on earlier versions of the article. DOI: 10.11588/huplc.2021.1.82302

ticular with respect to the way in which the "time spans" can be interpreted. What exactly is the time of the utterance in written language? And in a sentence such as *The cat was dead*, which is intuitively clearly in the past, does the "time of the situation" (the being dead of the cat) really precede the utterance time? We see that even in such a simple case it is not always possible to have a precise and intuitively correct definition. There are also many semantic concepts which cannot be easily stated in formal terms, for example *causation* or *possession*, or even *aspect*. Nevertheless, a formal point of departure can clearly serve in drawing up a viable *tertium comparationis*.

A further approach in the attempt to attain comparable semantic categories in cross-linguistic research is to present the speakers of different languages with the same task in which what they have to express is presented non-verbally; this can be a scene with a particular spatial constellation when studying spatial relations, or a video in the case of temporal structures. This elicitation procedure is well established in both first and second language acquisition studies (Berman and Slobin, 1994; Bowerman, 1996; Flecken et al., 2015; Hickmann and Hendriks, 2015), but also increasingly in language typology (Croft, 2003; Dancygier, 2017; Klein and Li, 2009; Krifka and Musan, 2012; Pederson and Bohnemeyer, 2010 to give but a few examples). It provides the researcher with rich material based on the same stimulus which allows a high degree of comparability at this level. It may not be perfect, because speakers sometimes interpret the task somewhat differently, but it brings the researcher closer to identifying what may hold as a similar underlying semantic concept and is thus suited to empirical studies which require reliable data bases. Nevertheless, one has to take into account that there is a limit to the semantic concepts that can be studied in this way. Elicitation procedures of this kind are applicable in domains such as time or space, but less so for causation, possession or counter factuality. In the research contexts to which they apply, however, they serve in providing the researcher with what may be termed "raw material" based on the actual responses of the speakers of the different languages studied, i.e., sentences or texts which must be analyzed.

A crucial first step in coding the material typically means that it has first to be segmented into smaller units. This is already a problem within and across standard languages. Would a sentence such as *Peter sieht ihn mit dem Nachbarn reden* 'Peter sees him talking to his neighbour' be coded as one unit? Would *Peter sieht, wie er mit dem Nachbarn redet* 'Peter sees, that he is talking to his neighbour' consist of two units? This relates to the problem of syntactic density – in this case an AcI–construction versus a subordinate clause construction – and the crucial question concerning the status of the syntactic format when coding the conceptual content. In the examples, both sentences refer to two situations in the external world. However, the situations in question are represented at different levels of integration. This raises the question as to how we can proceed given languages such as Chinese where use of morphological finiteness cannot serve as an indicator for semantic unit formation? This gets even more difficult with developing systems involving first or second language learners: *Friend street car hospital* – how does one code this text in terms of segmentation let alone aspectual categories? Linguistic data do not come neatly packaged.

When data analysis per language is completed, results have to be pooled and compared quantitatively. Although computer-assisted methods are available, both for coding as well as statistical evaluation, coding systems have to be developed which allow for both data exchange and cross-linguistic comparisons. There are a number of challenges which have to be met in this process, some of which are very difficult to solve. The most important ones concern the theoretical basis underlying the categories used in coding the data, the level of specificity selected and – first and foremost – their application in context to the concrete linguistic material. Given some of the problems indicated, the goal of this paper is to discuss solutions as well as the drawbacks with which we are confronted when coding multilingual data at the level of semantics.

2 Contexts for coding

The first lesson we learn when surveying semantic coding systems is the absence of a generally accepted coding schema in specific semantic fields. Rather, studies tend to proceed on the basis of individual coding schemas developed according to their specific goals. This stands in clear contrast to coding systems developed at the morphological or syntactic level of language. Without being naïve with respect to problems related to morpho-syntactic parsing and annotation, the fact that there are automatic tools which can be applied to these ends shows that there are, to some extent, sufficiently-defined categories which can be assigned to linguistic structures across languages. There are also attempts to develop se-

mantic parsers (cf. latest developments in this field Proceedings of the ACL Workshop, 2019). However, these tools operate on the linguistic surface form and can only capture certain aspects of meaning, as will be discussed below. Moreover, the application of a general set of values will run into problems when applied across typologically different languages, starting with standardization requirements which are not met by natural spoken language and learner languages and ending with the problem of semantic equivalence across languages.

In the following we will first outline the main dilemmas encountered in semantic coding in general. Then the analytic challenges will be illustrated for two semantic domains: space and time.

2.1 Three dilemmas

Why do we speak of dilemmas? Coders have to take numerous decisions with respect to different aspects of the data. Every decision which serves the function of achieving more clarity in one respect, does so at the expense of clarity in some other respect.

Dilemma 1

The first step in developing a coding system consists in identifying the unit of analysis. Semantic features can be coded at different levels, starting from the morpheme level as the elementary meaning-carrying-unit up to the discourse level including pragmatic components of a communicative exchange. Two requirements have to be balanced in cross-linguistic coding at the semantic level: limitation of factors potentially involved in the representation of meaning on the one hand, thereby taking into account, on the other hand, the maximal range of factors which could contribute to disambiguating meaning. While it might be useful to code meaning at the phrase level as categories can be assigned on the basis of general lexical knowledge, this could lead to misinterpretation given that principles of compositionality, coercion, and inferencing operate on the lexical/phrasal level and lead to reinterpretations. A simple example from the domain of definiteness will serve to clarify the problem:

(1) *The traitor was sentenced.*

(2) *The traitor was never found.*

(3) *The traitor is not a legal, but a moral category.*

The nominal phrase the traitor can be coded as definite at phrase level. Definite noun phrases in context can be associated with further features such as *givenness*, *specificity*, *identifiability*, *referentiality*. If the coding procedure takes the whole sentence into account then the stage is set for the inevitable discussion on the semantic-pragmatic status of definite NPs: What is identifiability, what is the status of NPs in generic sentences, etc. – a minefield for cross-linguistic research.

Dilemma 2

Closely connected to dilemma 1, we encounter the problem of explicitly conveyed and implicitly included meaning. The latter component has to be integrated, albeit with a different range. Coders could take the information at sentence level into account in order to disambiguate meaning, they could also take the preceding linguistic and even the situational context into account. The attempt to integrate inferred information when assigning semantic codes means the process gets fuzzier and less controllable. On the other hand, sticking closely to the literal verbal product may lead to incorrect or underspecified categorizations. Take the following example, in which spatial categories are at issue: *The man ran to the door. Standing outside he took a deep breath.* If we code for spatial information sentence by sentence then the first sentence will be categorized as a directed motion event with a potential endpoint, one which cannot be categorized as reached. If we take the text level as the basis in the coding process, then it is obvious that the goal *the door* has been reached – although this is not explicitly said. This piece of information will be inferred drawing automatically on several steps of default inferencing. This dilemma gets even more pronounced when we work on formally deviant data such as learner languages. Depending on the level of competence of the speaker, the interpretation of an utterance will hinge on inferencing to a large extent.

Dilemma 3

Another problem lies in the fact that terminology is normally defined in the context of specific semantic theories. There are numerous approaches to temporal or spatial semantics, for instance, a fact which often results in different definitions of the same term. This would not be a problem if the theoretical categories were not often tied to and illustrated by specific linguistic systems. The case of *progressivity* in the introduction makes this clear for categories of temporal aspect. If a categorization is developed on the

basis of the English system, then the category *progressive*, or more specifically *ongoingness*, includes those temporal features which constitute the meaning of the *ing* operator. If we include other aspectual systems in our analyses such as those in the Slavic or Arabic languages, then the category of *progressivity* describes a temporal operator which only partially overlaps with the English form in meaning. Again the dilemma lies, on the one hand, in the balancing act between being more general as basis for comparison or being more specific and therefore closer to the actual meaning expressed in a specific language.

There is no straightforward recipe in handling these dilemmas. A number of factors come into play. First and foremost, the respective research question should guide the decisions made with regard to coding. The research question presets, for the most part, in which semantic domain the given data should be coded at a level of high granularity, i.e. in a highly differentiated way, and where coding can be more general at the expense of semantic differentiation. This will be illustrated on the basis of case studies below.

Another important factor lies in the typological range. The coding schema has to be elaborated depending on the linguistic material under investigation. In the case of learner languages we might need more than one layer of semantic description, one providing the meaning of a verbal form in the target language, one giving an interpretation on the basis of all information available within the given common ground, as well as potentially post-hoc clarifications by the speaker. This again depends on the degree to which explicit information is lacking. The utterance *aber Mutter gesehen besser* 'but mother seen better' produced by a Turkish worker is ambiguous in different directions: who is the subject, who or what the object, what is the scope of *besser*? *Aber ich gesehen Mutter besser für mich* 'but I seen mother better for me' – in this case less of the message remains implicit. This means that coding for semantic properties can be more local and less speculative.

In the following we will illustrate these problems by looking at two semantic domains which are subject to innumerous studies both under typological as well as acquisitional perspectives. One covers the domain of spatial cognition with a focus on object location (4.1) and motion events (4.2), the other one concerns the domain of temporal perspective taking with a focus on aspectual properties of events (4.3).

3 Spatial categories

3.1 The case of static locations

Bowerman (1996) and Bowerman and Pederson (1992) studied spatial cognition under a typological as well as acquisitional perspective. In order to ensure comparability, subjects were shown pictures of objects in different spatial constellations. The task was to describe the situation, one object being marked as the figure to be localized. This methodology has set standards in that many follow up studies used these stimuli across a large number of languages and language varieties. Two examples below illustrate the type of stimuli.

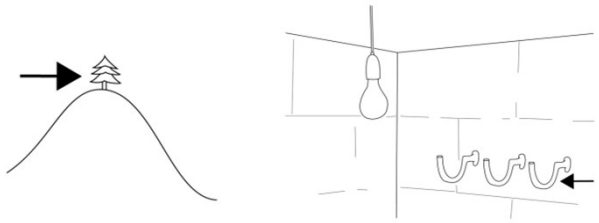

Figure 1: *Stimuli from the stimulus set Bowerman and Pederson (1992)*

The data obtained consist of utterances in the form of a nominal phrase and a locative *the towel hook on the wall* or a complete sentence *the towel hook is fixed to the wall/the towel hook on the wall.* The question is - what challenges are we faced with if we want to compare which spatial concepts are used across languages in solving this communicative task?

We will illustrate the coding task by looking at German, English and Turkish. The decisions which have to be made with respect to the coding categories relate both to the semantic features selected in capturing the meaning expressed in the different languages as well as the factual spatial relation between the objects in question. It is certainly not sufficient to simply use the linguistic expressions as a type of semantic meta-language, which is often the case, however, in cross-linguistic research - as in [*the tree on the hill* spatial relation: ON]. A more fine-grained description will draw on a bundle of semantic features, relevant in characterising and distinguishing the different spatial categories. Features in question could involve dimensionality (2/3), topology, specific dimension (vertical/horizontal/lateral), proximity (contact, margin, remote), functionality (support, attachment), frames of reference (deictic, intrinsic, absolute, mixed) (cf. Becker, 1994; Bowerman, 1996; Carroll, 1997; Croft, 2001; Feist, 2008;

Levinson, 1996, 2003; Pederson et al., 1998).

Given a comparative analysis coding can be organized along two different lines. The material can be structured according to the type of spatial relation shown in the stimuli: a) starting from the situation in the external world and asking the question how the languages express specific spatial constellations. b) starting from the expression and asking the question what situations are referred to by a specific expression. Both approaches are subject to problems which in the end require individual solutions. Figure 2 provide examples for each approach (see stimuli in figure 1).

Both approaches harbour problems. The first approach suggests that the clustering of situations based on the spatial relations between the objects in the real world is given by the factual nature of the situations. But speakers of different languages might cluster situations according to different criteria. For instance, the spatial concept of a *margin space*, the space adjoining the surface, is irrelevant for expressions in many languages (German in contrast to English and Turkish in the examples). Another case in point would be a situation in which one object is inside another and which could be seen as support in one language and containment in another (e.g. *Das Buch auf dem Tisch - the book on the table* (support), *el libro en las mesa* (containment)). This problem can only be solved pragmatically in that several coding cycles are passed through, going back to the data, refining and reorganizing the coding categories[1].

The second approach in which the linguistic data are coded for semantic features also requires considering the interplay between the development of the categorical grid and the diversity of the data obtained across languages. The semantic feature *loc in margin space*, for instance, has to be added to distinguish *an* from *auf* in German, a feature which is not relevant when describing the English or the Turkish system of basic locative expressions. Coding has to be as specific as necessary and as general as possible in answering the research question at hand. It can never grasp all aspects of variation in the domain of localizations which are manifested across languages. "Even if two languages appear on the surface to draw the same distinction, the boundaries between the contrasting categories often differ". (Feist, 2008, p. 1183)

If we now consider L2 data, the second procedure is not viable at the outset, in particular with early learner languages. Given the fact that the linguistic form as such does not provide sufficient information regarding the intended meaning, the first approach is required in this case. If the correlates in the external world are documented in controlled situations, then the organization of the data according to the conceptual space activated by the stimuli provides insights into form-function relations, potential L1 influence, as well as learner-specific over generalizations.

3.2 The case of motion events

Motion events are a central topic in cognitive and semantic typology (Carroll, 2000; Carroll et al., 2012; Flecken et al., 2015; Pederson et al., 1998; von Stutterheim et al., 2017; von Stutterheim et al., 2020). Talmy's seminal work laid the ground work for a whole field of research (Talmy, 1975, 2000). He suggested six basic conceptual/semantic categories in describing and contrasting the means used to express motion events across languages: *motion, path (source, trajectory, goal), figure, ground, manner* and *cause*. These categories have been adopted as a coding grid for linguistic data in innumerous studies. However, empirical research has shown that variation across languages cannot be sufficiently captured by using these features as the basis for comparison. It has become clear that languages differ not only with regard to the components of an external situation which are selected in encoding (e.g. path versus manner), they also differ with regard to how situations are segmented (e.g. break points given by changes at the level of the ground versus orientational changes by the figure, Gerwien and von Stutterheim, 2018) and the perspective selected when conceptualizing spatial relations (e.g. ground-based versus figure-based categories, Carroll et al., 2012), the relation between implicit and explicit information, the extent to which interactions with other conceptual domains are systematically taken as contributing to the spatial meaning of an utterance (e.g. interaction with temporal categories, von Stutterheim et al., 2017) to name the most prominent differences. Operating at this level of differentiation reveals that languages, which according to Talmy's typology belong to one category, actually have to be placed in separate typological (sub)boxes. The integration of these layers when coding and thereby going beyond the basic notional categories poses enormous challenges for

[1] The extensive study by Levinson (2003), can be taken as a model when integrating linguistic data from highly varying languages. Specific semantic properties of spatial expressions which are not categorized as part of a pool of more widely shared notions are integrated by a special row of 'notes' (p. 439).

Type A

Relation between objects on picture	Language	Expression
Located at surface/contact, support	German	Der Baum auf dem Gipfel. (the tree on the hill) 'The tree *on (top of)* the hill'
	Turkish	Dağ*da* ağaç (*hill LOC tree*) 'The tree *on (top of)* the hill'
Located at surface/contact, support	German	Der Haken *an* der Wand (*the hook at*-on *the wall*) 'The hook on the wall'
	Turkish	Kancalar duvarda (hook wall LOC) 'The hook on the wall'

Type B

+ = semantic feature of spatial expression, - = semantic feature is excluded, o = semantic feature is not represented

Expression	Language	Contact	loc in margin space	support	loc relation
Der Baum *auf* dem Gipfel	German	+	-	+	o
Der Haken *an* der Wand	German	o	+	+/-	o
The tree *on (top of)* the summit	English	+	-	+	o
The hook *on* the wall	English	+	-	+	o
Dağ*da* ağaç	Turkish	o	o	o	+
Kancalar duvar*da*	Turkish	o	o	o	+

Figure 2: *Example tables for the two coding types*

comparative studies. We will illustrate problems and solutions on the basis of a current study on motion event segmentation in French, German, and English. In the study in question three groups of participants with three different language backgrounds were asked to describe short video clips. The stimuli showed two types of motion events. In one group of stimuli a figure moved continuously along a path in the course of which the figure changed orientation/direction. The second group depicted a figure moving along a continuous path with no change in direction or orientation.

Figure 3: *Scene: A woman walking around a fountain and up some steps*

The research questions addressed on the basis of the data obtained related (a) to typological differences in event unit formation, (b) the information selected in representing segments of the scene and (c) the distribution of information across formal categories. In order to pursue these questions, the data – which was collected in different countries and by different researchers – had to be coded in a standardised way. Data preparation thus entailed the following number of steps.

Step 1: Cleansing

In this case it has to be decided whether, and if so, which parts of the data are not relevant given the research question and should thus be excluded. In the case in question, utterances which did not refer to the motion event as such (e.g. descriptions of figures, objects and background or evaluations and comments by the speaker) were not considered in

Figure 4: *Scene: A woman walking uphill*

Figure 5: *Scene: a man is walking along a street, he turns left and enters a large building through an archway (2 path segments)*

the analyses, for example, utterances such as *apparently the person was searching for something or the man disappeared from the screen.* Less clear was the criterion for syntactic packaging. While it was to be expected that subjects would produce full sentences, some used only nominal phrases or verb-less constructions when referring to specific scenes. The question is should these utterances be included in a comparative study on event construal as events are dynamic, referring to a bounded temporal interval? This type of information is typically expressed by verbs in the languages included in the study. But what is the difference between *un homme sur un vélo* 'a man on bicycle' and *un homme fait du vélo* 'a man does of bicycle', given that they both refer to a video clip in which a man is cycling down a street? The decision can be based on different criteria. First there is the criterion of relevance: If this is a single or a very rare case then it should be excluded. Then there is the criterion of design artefact: Is the particular form selected by subjects confined to a particular item, and if so, then this points to a problem at the level of stimulus design. Finally, there is the criterion of individually divergent strategies: If there is an individual speaker who chooses an encoding strategy across all items, e.g. just naming the figure, then this data set should be excluded.

Step 2: Selection and definition of the coding categories

The following examples illustrate that Talmy's categories are sometimes not sufficiently fine-grained to capture the relevant language specific principles. The examples listed are from the French, German and English corpus.

(4) *Un homme marche dans la rue et rentre dans un bâtiment.*
'a man walks in the street and enters in a building'

(5) *Un homme marche et rentre dans un bâtiment.*
'a man walks and enters in a building'

(6) *Ein Mann geht in ein Haus.*
'a man walks in a house'

(7) *A man is walking along a street, he turns left and enters a building.*

In the light of the three research questions introduced above the semantic coding schema has to cover the following categories in order to represent the patterns in the three languages.

a) number of path segments referred to by the speakers

b) specific path segments selected

c) spatial concepts selected for encoding the respective segments: location, ground (source, trajectory, goal), reference frame (figure oriented, deictic), manner

d) packaging in syntactic categories

Step 3: Applying the codes to the data

With this step the coder is confronted with a number of challenges which will be discussed using the French, German, and English examples under (4) to (7) for illustration.

Number of segments

What seems to be fairly straightforward turns out to be problematic. The first French example ((4)) can be categorized as containing two path segments. These are expressed by the two spatial adjuncts *dans*

la rue 'in the street' (segment 1) *et dans un bâtiment* 'in a building' (segment 2). The two phrases refer to different spatial segments of the path, when the concept *path* is described on the basis of features of the ground. The second example in French ((5)) describes the same situation. However, it is not the case that the two sentences refer to two segments of a path. The figure does not change its manner of motion - he is walking all the way. Rather the first sentence could be interpreted as not referring to the 'path' at all, since the concept of *direction* is not involved, a feature which is typically associated with *path*. One can walk around in circles without going anywhere. In this sense the first sentence in (5) simply expresses manner of motion and only the second sentence provides information on the path taken. In this case one would code for only one segment.

Path segments selected

Here, the French example does not pose any problems once a decision has been taken on how to treat bare manner verbs. The description in English, however, raises the problem as to how to categorize the *turning*-event. Is it a segment in its own right or is it a break point between segments? The case of *to turn* in example (7) illustrates a problem which occurs if there is no language external control of the content expressed. A *turning*-event can have different spatial properties depending on whether it refers to a change of position of a figure or a change in directed motion: *The man turned around* versus *the car turned around the corner*. The solution will depend on how the concept of 'path' is theoretically defined in the analysis, as indicated above. The path can relate exclusively to contours of the ground, or the trajectory drawn by the figure in motion, or both. For example, if the point at which the figure in motion is described as *turns left* is located on a path such as a street which continues without any change in direction, the change taken when 'turning left' is defined by the change in orientation of the figure in motion. In this case there are three path segments: 'goes along x', 'turns', 'enters y'. If the features of the ground are the point of reference then there are two paths: moving "along the street", and the transition in moving "from the street into the building". In German there are two different verbs which have to be used in these cases: *sich umdrehen/drehen* 'turn around' for the change in position and *ab-/einbiegen* 'turn into' for the change in direction along a path. Unless we have details on relevant features of the external situation, we cannot decide whether the second sentence in the English example refers to a path segment, which remains unexpressed in the verbalisations in the other languages, or whether it describes a change in the orientation of the figure which does not extend through space.

Spatial concepts selected

In many, but not in all cases, Talmy's categories are sufficient. Consider the following descriptions in German and French.

(8) *Un homme* (figure) *marche* (manner) *dans la rue* (ground/location) *et rentre* (path/directed motion) *dans un bâtiment* (place at goal).
'a man walks and enters a building'

(9) *Ein Mann* (figure) *geht* (manner) *in ein Haus* (goal/directional).
'a man walks in a house'

The coding shows a difference which is well described for German and French. Directionality is expressed by the path verb (*rentrer* 'to return') in French, while it is expressed by the accusative case on the adjunct in German. What the coding does not show, however, is the fact that German and French manner verbs behave differently in semantic terms. In German, manner verbs encode manner and directed motion, they are dynamic. This is evidenced by the fact that they have to combine with a spatial adjunct when used in the context of a concrete motion event. *Ein Mann läuft* 'a man walks' can only be used if the speaker intends to contrast manner of motion of a specific type with another form of motion, e.g. the man doesn't drive. In this case it is not a motion event which is expressed. In French, manner verbs behave differently: the sentence *un homme marche* 'a man walks' focuses on a property of the figure rather than a motion event with displacement (von Stutterheim and Gerwien, 2021). This example shows that well established semantic categories, as typically used when coding cross-linguistically, may mask differences which can only be identified given in depth analyses of language in context.

Step 4: The adaptation cycle - adapting coding procedures inductively on the basis of the data

The considerations outlined above already point to the next important step required. The integration of more data showing further ranges of variation in

one language, as well as data from other languages, requires what may amount to a series of re-analysis of the coding categories with the necessity of additional categories and/or refinement of those already applied. This will be illustrated with examples from English and French.

(10) *A man* (figure) *is walking* (manner) *along the street* (ground/path)

(11) *Un homme* (figure) *s'approche* (path/directed motion) *d'un bâtiment* (ground/goal location)
'a man approaches of a building'

Focusing on path, this coding procedure suggests that this concept is expressed in different syntactic categories, with no difference at the level of semantics. This coding pattern can be found in numerous typological studies on motion events (e.g. Bohnemeyer et al., 2007; Cifuentes-Férez and Gentner, 2006). However, there is a conceptual difference between the spatial concepts selected in representing the path of motion. In French, the spatial concept is derived from the orientation of the figure, with no information on the possible role of properties of the ground in specifying features of the path; in English, by contrast, the spatial concepts are derived from features of the ground (cf. Carroll, 2000; Carroll et al., 2012). These conceptual contrasts are crucial in understanding the constraints found in languages such as French (dominant spatial concept: figure - based) in contrast to English or German (dominant spatial concept: ground - based). Given these findings, the coding schema would have to be adapted by adding a category for figure- vs. ground-based concepts.

In the discussion so far, considerations have not been extended beyond the spatial domain. However, studies have shown how spatial concepts closely interact with temporal categories, aspectual marking in particular, in the expression of motion events (Bylund et al., 2013; von Stutterheim et al., 2012). This factor would then require the inclusion of temporal categories in the coding schema. Problems related to the conceptual domain of temporality in more general terms will be illustrated in the following section, which also takes a look at learner languages.

4 Temporal categories

A study on aspectual marking in Chinese, with comparisons involving German (first language = L1) and very advanced Chinese learners of German (second language = L2) serves to illustrate the challenges and possible solutions when coding temporal categories. The data base consists of oral event descriptions which were elicited based on short real-world video clips as stimuli. The research question addressed the topic of language specificity in aspect marking and the question of possible traces of conceptual transfer from L1 to L2. Given this context and aims of the study, the semantic categories assigned in all three sets of data had to be equivalents of the same temporal concepts. A number of examples will serve as a starting point to illustrate the steps which have to be taken in order obtain a valid coding schema. The scene described by the subjects below shows *a woman cleaning a table*, and closes with the woman walking away.

German L1

(12) *Eine Frau wischt einen Tisch ab.*
'a woman wipes a table off.'

German L2 with Chinese L1

(13) *Jemand putzt einen Tisch fertig.*
'somebody cleans a table done.'

(14) *Eine Frau putzte den Tisch.*
'a woman cleaned the table'

(15) *Eine Frau hat den Tisch geputzt.*
'a woman has the table cleaned'

Chinese L1

(16) *mou3 ren2 zai4 ca1 zhuo1zi..*
'somebody ZAI-marker wipe table'

(17) *you(3) ren(2) ca(1) wan(2) le zhe(4) zhang(1) zhuo(1) zi..*
'somebody wipe complete LE-marker this table'

Step 1 Cleansing

Typically, there are many aspects of the data which are not relevant for the question at hand and should therefore be excluded. In this case, the relevant data consists of event descriptions of the scene shown in the video; the semantic features under focus in the analysis are temporal, or more specifically, they concern categories which express aspectual distinctions. Apart from data which is technically defective, the data excluded will include subjects who produce only a noun phrase or some other type of sentence

structure which is incomplete. If a person verbalizes more than one sentence then it has to be decided whether it is necessary to include all sentences or select the one referring to the event shown in the clip (cf. the discussion of coding data in L1 acquisition in Bassano and Hickmann, 2013, p.111). The decision depends again on the specific aims of the study. If the question relates to the way speakers of different languages select and structure information given a specific visual input, a question anchored in cognitive linguistics, then all linguistic material produced has to be taken into account. If, however, the question relates to aspectual perspective taking in event descriptions, then the units selected can be confined to those relating to events.

Step 2 Selection and definition of the coding categories

Given the case where the aspectual options induced by the stimuli are restricted to selecting different phases out of a concrete event, this will exclude, for this specific data set, habitual or generic aspectual categories. The languages involved in the study in question are Chinese, German and the associated learner language of German. Although it is not excluded that learners may produce something which is completely alien in either their L1 or L2, they would not belong in the categories set out in the first step in the analyses. With regard to the relevant contrasts for the learners, Standard German does not feature aspectual morphology: Verbs are inflected for tense but not for aspect, while Chinese expresses aspectual categories by means of particles. Although these forms are to some extent grammaticalized, they are not obligatory. The relevant forms for the data in hand are the particles *le*, usually described as a perfective marker, and *zai*, which is described as marking *ongoingness*. This is not sufficient as such, however, since a theoretical framework has to be adopted in order to clearly specify the temporal coding categories at issue. For purposes of illustration, the coding schema will be based on Klein's theory on time in language. Given the research question outlined above, the following features can be coded at the outset as follows (Klein, 1994; Klein, 2010):

- the inherent temporal properties of the lexical material used, resulting in 1-state or 2-state situations (*inherent aspect* in the following)
- the aspectual relation between topic time[2] (TT) and situation time (TS) with 5 subcategories
  - TT includes TS
  - TT is after TS
  - TS includes TT
  - TT precedes TS
  - TT includes parts of TS and of post state of TS

Step 3 Applying the codes to the data

Following the same procedures as outlined above with data in the spatial domain, the same sentences will be used to illustrate the coding process.

(18) a) *Eine Frau wischt einen Tisch ab.*
'a woman wipes a table off'
inherent aspect: einen Tisch (abwischen – 2-state predicate) 2-state predicate
TT – TS: TT includes TS

(19) a) *Jemand putzt einen Tisch fertig.*
'somebody cleans a table done'
inherent aspect: (einen Tisch (fertig (putzen – 1-state predicate) 2-state predicate)2-sp)
TT – TS: TT includes TS
TT – TS: TS includes TT

b) *Eine Frau hat den Tisch geputzt.*
'a woman has the table cleaned'
inherent aspect: (einen Tisch (putzen – 1 -state predicate) 1-state predicate)
?TT – TS: TT after TS

c) *Eine Frau putzte den Tisch.*
'a woman cleaned the table'
inherent aspect: (einen Tisch (putzen – 1-state predicate) 1-state predicate)
?TT – TS: TT includes TS

(20) a) *mou3 ren2 zai4 ca1 zhuo1zi.*
'somebody ZAI-marker wipe table'
inherent aspect a table (clean–1-state predicate) 1-state predicate)
TT-TS: TS includes TT

b) *you(3) ren(2) ca(1) wan(2) le zhe(4) zhang(1) zhuo(1).*

[2] Time of situation is the time interval which the situation described by the lexical content takes, topic time is the temporal interval about which a claim is made (cf. Klein, 1994)

'somebody wipe complete LE-marker this table'
inherent aspect a table (clean–1-state predicate) 1-state predicate)
TT-TS: TT in post time of TS

One could list a number of objections to the way this small corpus is coded for aspectual features. We will begin with the German example. In example (18 a) the inherent temporal properties cannot be disputed. *Abwischen* 'to wipe off' is a 2-state-verb. However, the present tense raises an ambiguity with respect to the TT-TS relation. It can either express *ongoingness* or it can refer to the event holistically. The latter option is selected in the coding. However, the linguistic material as such cannot disambiguate between these two perspectives. The correlate in the external world is also problematic. The subjects who were presented with two variants of the same scene in the course of the experiment – one in which the action was carried out until a resultant state was reached, one in which the action was not fully carried out – used the same wording in German L1. How can the coder cope with this problem? We suggest to code as 'category not specified' in these cases.

The learner data pose problems of a different nature. On the one hand, the aspectual perspective is explicitly expressed by a (non-native) choice *fertig* 'done'. On the other hand, however, the German verbal tenses are used in a way not found in the data of the native speakers. The present perfect in German (19 b) has two readings: It can either refer to a situation in the past, exchangeable with the simple past or it can refer to the post state of a situation, as with the English present perfect. What the L2 speaker actually has in mind when using this form is not clear. Since Chinese native speakers tend to mark perfective aspect in relation to these scenes, the coder might be tempted to code for perfective aspect. But this is not unambiguously backed by the data. The simple past used in the last L2-sentence (see (19 c)) is even harder to categorize: The form is not adequate in the given context from an L1 point of view. So does the learner use this form in order to find a way to express perfective aspect, i.e. a compensatory strategy? Or does he choose an idiosyncratic strategy with the actions located in the past? Or does he simply mispronounce the form slightly? Again, this cannot be decided on the basis of one sentence. In these cases it is helpful for the coder to check the entire data set which may allow the identification of individual learner strategies as well as specific form-function relations.

The Chinese L1-data in (20) do not pose severe problems – at first sight at least. Although the explicit aspectual markers *zai* and *le*, as aspectual categories can be annotated in a straightforward manner, what remains unclear, however, is the choice of different aspects in relation to the same external situation and the same communicative context. The aspectual markers are not obligatory in Chinese. While English speaking subjects all select the *–ing* form (100) in this context (von Stutterheim et al., 2012), Chinese speakers do not show a consistent picture. This points to a question of categorization which arises given the range of languages which have to be integrated in the analysis.

Step 4 The adaptation cycle - adapting the codes inductively on the basis of the data

If we include other languages with grammaticalised aspectual systems, the categories introduced above will clearly not suffice. The *–ing* in English cannot be placed in the same aspectual category as *zai*; the same would apply to the Russian imperfective if Russian were included in the corpus: the *imperfectivity* or *ongoingness* expressed by these means are similar but they are not the same in the sense that all relevant factors do not overlap. The adaptation cycle has to start with data refinement, along with a relevant overhaul of the coding categories, so that the specific features required in answering the respective research question are brought into focus.

5 Lessons to be learned

The cases examined above are intended to bring certain problems and challenges to the fore. It should be emphasized that they are by no means exhaustive discussions of coding systems for the temporal and spatial categories selected for illustration. The general conclusions drawn from these considerations relate, on the one hand, to questions concerning methodology and the toolbox used, and on another level to the theoretical issues concerned. As a starting point we will pinpoint the major challenges which have emerged in the discussion of the three cases above.

With regard to theory-based coding categories such as those outlined above, the crucial challenge consists in developing categories which (a) are sensitive to language contrasts without being trapped in a language-specific bias and which (b) are sufficiently

fine grained in capturing those aspects of the semantics of the data which are relevant in answering the respective research question.

As regards the coding procedure, the crucial challenge consists in coping with the discrepancy between what is *said* and what is *meant*. Deficient or ambiguous linguistic material is the rule and not the exception in the language production of learners as well as standard languages.

So how can these challenges be met? We will conclude with a number of suggestions which should be taken into consideration by coders in the context of semantic typological studies.

(a) The general coding strategy has to be oriented towards the **specific research question**. Since coding is extremely time-consuming, the level of granularity should be selectively adapted in terms defined by the aims of the study. This could mean that data may be coded at a high level of specificity in one domain, e.g. in the spatial domain, with the remainder represented at a general level. If the goal of a study is to represent features of as many languages as possible, as in the case of WALS[3] , for example, then feature assignment has to be broad. The explanation given under the category aspect in WALS illustrates this very clearly:

"We distinguish **imperfectives** from progressives, with which they partially overlap and which are often seen as a variety of imperfectives. Progressives, as the English is singing or the equivalent Spanish está cantando, have a more restricted domain of use (for instance, they are typically not the primary choice for expressing habitual meaning), which means that they are opposed to non-progressive forms independently of time reference. They are also normally restricted to non-stative verbs. Progressives are frequent diachronic sources of marked imperfectives, and borderline cases admittedly exist. ..... For this map only two values have been defined: languages in which there is grammatical marking of the perfective/imperfective distinction and those where there is not." (http://wals.info/chapter/65)

The fine grained difference in the aspectual systems mentioned in the commentary may be highly relevant in a study on psycholinguistic questions in the context of spatial cognitive typology.

(b) In order to ensure valid coding categories, the particular theoretical approach is not that important. But one way or the other the categories must be anchored in **a well - defined semantic theory**. This constitutes the only basis for developing a common categorical framework – a factor which is particularly challenging in typological studies with researchers from different linguistic traditions in their own languages. As anyone working on semantic coding across different languages can claim, one pays later for omissions at this level.

(c) In the course of implementing the codes the following aspects have to be borne in mind. **Data preparation should be kept as controlled as possible**. If the documentation of the occurrence and use of specific forms is at issue, then large corpora are required. However, if the research question relates to cognitive correlates of certain linguistic forms, then unscripted speech presents a problematic data base. A high level of control over the content of linguistic data used, means that the 'meant' can be better taken into account[4].

(d) When preparing the coding it is necessary to **invest sufficient time** in a common inventory and understanding of the coding categories. Given the initial procedures in setting up the categories under consideration, the system has then to be elaborated on the basis of the data at issue. This could mean becoming more fine grained in the critical domains, or that other domains have to be integrated which had not been taken into account in the initial stage (e.g. temporal categories analyzed in the context of spatial factors in the domain of reference); it could also mean that language specific categories have to be developed which can only be integrated into a common framework at the level of data analysis (cf. aspectual distinctions).

(e) In applying the coding **be prepared to rework** the coding system where necessary. In this regard, decisions have to be taken which may not necessarily fit the theory in obvious terms. This starts with the definition of the units of analysis. The multifaceted options of packaging information, whose status is less clear syntactically, for example, requires agreement across researchers on how to treat the different forms in the coding system (e.g. non-finite forms or sentences with no verb). The second

[3] Dryer and Haspelmath (eds.), 2013

[4] Good examples for typological research addressing acquisitional and psycholinguistic questions are the studies by Berman and Slobin (Berman and Slobin, 1994) including many follow-up studies on a wide range of languages using the picture book Frog, where are you, the studies by Hickmann and Hendriks using the picture stories Horse and cat story (Hickmann and Hendriks, 2015).

challenge relates to the discrepancy between 'what is *meant*' and 'what is *said*'. It is absolutely vital that coders concur in the range of contextual knowledge which is taken into account when assigning semantic features. In data involving language acquisition it may be advisable to code semantic categories at different levels, close to 'what is *said*' and additionally interpreted on the basis of contextual information. In general, the coding strategy has to be based on the principle of 'faithful to the data'. Where problems arise with respect to the categorization of the data, it is advisable to leave coding slots unspecified rather than force the linguistic material into categories which do not fit. It goes without saying that in many cases the solution may require a new elicitation procedure for a subset of the categories in order to gain a reliable picture of the factors at issue.

6 Concluding remark

There is no golden rule when coding semantic categories across languages. In order to cope with the dilemmas outlined above, researchers have to develop specific solutions on the basis of the research questions at issue and the languages involved in the study. A certain degree of subjectivity will, however, always remain when coding data at a semantic level. In addition to questions concerning ambiguity and under-specification, there is the problem of a language specific bias and its transferal by the coder to data from other languages. The costs that this may incur in distorting the analysis, in the sense outlined above, can be reduced by in-depth discussions between coders which may often require revisions over a series of coding cycles. In contrast to coding procedures at other levels of linguistic data, we are still far from being in a position to draw on automatic computer aided systems which go beyond meaning attributions based on forms at the level of words and sentences.

References

(n.d.). http : / / datavyu . org / user - guide / best - practices / tutorials / develop - a - coding - manual.html accessed 2017-06-27

Bassano, D., & Hickmann, M. (Eds.). (2013). *Grammaticalization and first language acquisition: Crosslinguistic perspectives*. John Benjamins.

Becker, A. (1994). *Lokalisierungsausdrücke im Sprachvergleich*. Niemeyer.

Berman, R., & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Erlbaum.

Bohnemeyer, J., Ibarretxe-Antuñano, I., Enfield, N. J., Kita, S., Ameka, F. K., Essegbey, J., & Lüpke, F. (2007). Principles of event segmentation in language: The case of motion events. *Language*, *83*, 495–532. https://doi.org/10.1353/lan.2007.0116

Bowerman, M. (1996). Learning how to structure space for language: A crosslinguistic perspective. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garret (Eds.), *Language and space* (pp. 385–436). The MIT Press.

Bowerman, M., & Pederson, E. (1992). Topological relations picture series. In S. C. Levinson (Ed.). Max Planck Institute for Psycholinguistics. https://doi.org/10.17617/2.883589

Bylund, E., Athanasopoulos, P., & Oostendorp, M. (2013). Motion event cognition and grammatical aspect: Evidence from afrikaans. *Linguistics*, *51*, 929–955. https://doi.org/10.1515/ling-2013-0033

Carroll, M. (1997). Changing place in english and german: Language specific preferences in the conceptualization of spatial relations. In J. Nuyts & E. Pederson (Eds.), *Language and conceptualization* (pp. 137–161). Cambridge University Press.

Carroll, M. (2000). *Representing path in language production in english and german: Alternative perspective on figure and ground* (C. Habel & C. von Stutterheim, Eds.). Niemeyer.

Carroll, M., Weimar, K., Flecken, M., Lambert, M., & von Stutterheim, C. (2012). Tracing trajectories: Motion event construal by advanced l2 french-english and l2 french-german speakers. *Language, Interaction and AcquisitionLanguage, Interaction and Acquisition. Langage, Interaction et Acquisition*, *3*, 202–230. https://doi.org/10.1075/lia.3.2.03car

Cifuentes-Férez, P., & Gentner, D. (2006). Naming motion events in spanish and english. *Cognitive Linguistics*, *17*, 443–462. https://doi.org/10.1515/COG.2006.014

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Croft, W. (2003). *Typology and universals* (2nd ed.). Cambridge University Press.

Dancygier, B. (Ed.). (2017). *The cambridge handbook of cognitive linguistics (part v: Methodological approaches)*. Cambridge University Press.

Dryer, M. S., & Haspelmath (eds.), M. ((2013). *The world atlas of language structures online.* http://wals.info accessed 2017-06-25

Feist, M. I. (2008). Space between languages. *Cognitive Science*, *32*, 1177–1199. https://doi.org/10.1080/03640210802152335

Flecken, M., Carroll, M., Weimar, K., & von Stutterheim, C. (2015). Driving along the road or heading for the village? conceptual differences underlying motion event encoding in french, german, and french-german l2 users. *The Modern Language Journal*, *99*, 100–122. https://doi.org/10.1111/j.1540-4781.2015.12181.x

Gerwien, J., & von Stutterheim, C. (2018). Event segmentation: Cross-linguistic differences in verbal and non-verbal tasks. *Cognition*, *180*, 225–237. https://doi.org/10.1016/j.cognition.2018.07.008

Hickmann, M., & Hendriks, H. (2015). Time talk in narrative discourse: Evidence from child and adult language acquisition. In J. Guéron (Ed.), *Sentence and discourse* (pp. 93–120). Oxford University Press.

Klein, W. (1994). *Time in language*. Routledge.

Klein, W. (2010). On times and arguments. *Linguistics*, *48*, 1221–1253. https://doi.org/10.1515/LING.2010.040

Klein, W., & Li, P. (Eds.). (2009). *The expression of time*. De Gruyter.

Krifka, M., & Musan, R. (2012). Information structure: Overview and linguistic issues. In M. Krifka & R. Musan (Eds.), *The expression of information structure*. De Gruyter.

Levinson, S. C. (1996). Language and space. *Annual Review of Anthropology*, *25*, 353–382. https://doi.org/10.1146/annurev.anthro.25.1.353

Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity*. Cambridge University Press.

Pederson, E., & Bohnemeyer, J. (2010). On representing events - an introduction. In J. Bohnemeyer & E. Pederson (Eds.), *Event representation in language and cognition*. Cambridge University Press.

Pederson, E., Danziger, E., Wilkins, D., Levinson, S., Kita, S., & Senft, G. (1998). Semantic typology and spatial conceptualization. *Language*, *74*, 557–589. https://doi.org/10.1353/lan.1998.0074

Proceedings of the ACL Workshop. (2019). *The first international workshop on designing meaning representations*. https://www.aclweb.org/anthology/W19-33.pdf

Talmy, L. (1975). *Semantics and syntax of motion*. Brill.

Talmy, L. (2000). *Towards a cognitive semantics vol.2: Typology and process in concept structuring*. MIT press.

von Stutterheim, C., Andermann, M., Carroll, M., Flecken, M., & Schmiedtová, B. (2012). How grammaticized concepts shape event conceptualization in language production: Insights from linguistic analysis, eye tracking data, and memory performance. *Linguistics*, *50*, 833–867. https://doi.org/10.1515/ling-2012-0026

von Stutterheim, C., Bouhaous, A., & Carroll, M. (2017). From time to space: The impact of aspectual categories on the construal of motion events: The case of tunisian arabic and modern standard arabic. *Linguistics*, *55*, 207–249. https://doi.org/10.1515/ling-2016-0038

von Stutterheim, C., & Gerwien, J. (2021). Die Bedeutung sprachspezifischer Ereignisschemata für die Argumentstruktur: Ein Vergleich zwischen dem Ausdruck von Bewegungsereignissen im Deutschen und im Französischen. In J. M. Hartmann & A. Wöllstein (Eds.), *Propositionale Argumente im Sprachvergleich: Theorie und Empirie* (pp. 385–436). Narr Francke Attempto.

von Stutterheim, C., Gerwien, J., Bouhaous, A., Carroll, M., & Lambert, M. (2020). What makes up a reportable event in a language? motion events as an important test domain in linguistic typology. *Linguistics*, *58*.