Building a Corpus in Linguistic Anthropology, the Example of Thakbong

Oliver Streiter
National University of Kaohsiung

Yoann Goudin
Ecole des Hautes Etudes en Sciences Sociales, Paris

Ann Mei-fang Lin
National University of Kaohsiung
Abstract:

In anthropology and linguistics, the term corpus refers to a collection of data sets used to tackle a particular research question. This can be a collection of passive sentences used for a research on the passive of a certain language, or a collection of photos, interview recordings and their transcriptions used for the analysis of a rite in a society. We will call such this a private corpus, which is designed for personal usage mainly. Another meaning of the term corpus, as used in Corpus Linguistics (Biber et al. 1998) and Computational Linguistics, refers to a systematic collection of data sets that are sampled with the aim to represent, as accurately as possible, a linguistic reality. We will call such a collection a public corpus. In this article, we describe how the notion of a public corpus can be extended to the research field of linguistic anthropology, first at the level of a concrete corpus-building project, and then, on a generalized level, covering a potentially wide range of possible research fields. We will first distinguish and discuss the levels of such a corpus, i.e. media, indices, transcriptions, descriptions, annotations and references. Then we will propose measures to evaluate the representability of the corpus in contexts where random sampling is difficult or impossible. Before doing so, we will define the term linguistic anthropology as understood in this article, and then, develop a research question from which a corpus-building project is derived.

Key words: corpus, Linguistic, Anthropology, Linguistic Anthropology, Corpus Linguistic, Thakbong, identity
1. Introduction: Two types of corpora

1.1. Private Corpora

In anthropology and linguistics, the term corpus refers to a collection of data sets used to tackle a particular research question. This can be a collection of passive sentences used for research on the passive of a language, or a collection of photos, interview recordings and their transcriptions used for the analysis of a rite.

Such a corpus is frequently constructed to be used exclusively by the researcher who collected the data. It should not be used or is not used by other researchers, as people lack the motivation or trust to share their data:

"One of the most revealing events of the conference came at the nightly presentation of slides of funerary and grave ritual taken by anthropologists at their field sites. It was startling (...) to discover that the six anthropologists present, who had been acquainted with one another for many years, had not previously seen each other's slides."

(Rawski 1988, pg. 21)

Designed for personal usage mainly, such corpora cannot be shared if not from the beginning designed for this purpose, as they are likely to be poorly structured and poorly annotated. For the individual researcher who remembers the circumstances of the data collection, a painstakingly annotation of 'obvious' information seems to be an unnecessary burden and is reduced to an absolute minimum, possibly in idiosyncratic abbreviations.

As a consequence, a private corpus is not suited for generalizations as features representing similarities, times and places are not made explicit: "He [i.e. Maurice Freedman] criticized recent work in Chinese anthropology for being too close to its informants and thus having a fragmented, peasant's eye view of
Chinese religion that emphasized diversity and variation" (Naqin, 1988, pg. 37). In order to be able to obtain generalizations from private corpora, corpora might be stuffed with authoritative voices. We find this tendency in Linguistics and in Anthropology.

Linguists look frequently at printed examples of standardized languages and then generalize to millions of speakers, without checking variations within the linguistic behavior covered by that language name. Actually, more often than not, the origin of the language data is not explained at all, as if it would not be important to explain whether at all and if so where and when and by whom and in which context something has been said:

"Data in (3) come from standard English. Data in (4) come from West Ulster Irish English."

(Boeckx 2007, pg. 36)

The authors even don't explain what they understand by the language name they use. Ethnologue (http://www.ethnologue.com), for example, mentions, among others, British Standard English, Australian Standard English, a Standard English that forms a continuum with the Guyanese Creole English, a Standard English that forms a continuum with the Leeward Caribbean Creole English, a Standard English and a Near Standard English in Hawaii and it does not mention an East Ulster Irish English. Thus using the reifying term 'Standard English', the sentence is claimed to be representative for an actually undefined range of language varieties.

There is an urgent need to specify such basic data, look at more difficult cases. Coupe (2007) calls the language of his analysis Mongsen Ao, spoken in Naga, Ethnologue lists a language with the name Naga Ao that has a dialect called Mongsen Khari. Is Mongson Ao thus the same as Naga Ao or as Mongsen Khari? A name of a language as classification for a linguistic example is thus very little informative and should not be used without data on time, place and speakers. In other words the language name is already a generalization and generalizations should be derived from data, not imposed on data. A private corpus, thus, is scientifically far from being basis for research.
The usage of written, authoritative resources for anthropological research has the same purpose, to obtain a seemingly valid, but basically unvalidated, generalization from reified sources:

"He could easily have taken Chinese historians to task (as he did one generation of sociologists) for a similar dependence on their sources and a resulting preference for seeing China through the unifying vision of its elites, a vision in which differences were minimized and deviations deplored."

(Naqin, 1988, pg.37)

1.2. Public Corpora

Another meaning of the term *corpus*, as used in Corpus Linguistics (Biber et al. 1998) and Computational Linguistics, refers to a systematic collection of data sets that are sampled with the aim to represent, as accurately as possible, a linguistic reality. We will call such a collection a **public corpus**. The success of public corpora is due to the fact that the criteria for the construction of the corpus go beyond the requirements for a specific piece of research and thus allow ever new research questions to be tackled through public corpora. Such corpora are generally distributed within the research community and cannot rely on implicit information.

A possible difficulty with public corpora is defining the reality to be represented, or sampling during the corpus construction such, that it is not representative of the reality. The indefinable term *balanced corpus*, c.f. Biber et al. (1998) is used in corpus linguistics to camouflage the lack of a definition of the reality of a 'language' to be represented. More than a workaround, but a change in theoretical stance, is to focus on 'languaging' (Foley1997) in the form of a collection of individual speech events, street signs, advertisements, publications, radio transmissions etc. Such types of languaging are much more easy to sample, as the population from which is sampled can be known or estimated. However such sampling will not go unchallenged with respect to its theoretical stance, since this approach, remarkably, renounces to the claim to sample a 'language' or a 'culture' and implicitly or explicitly threatens the uniformity
assumption of 'language' and 'culture' which underlies national discourses and the scientific discourses rooted in it. If a language or a culture cannot be sampled, these notions are denied the status of real world objects. As a consequence, these notions exist only in the discourse and as such become questionable.

Notwithstanding the difficulties in sampling, public corpora gain in popularity and ever new types of corpora are being developed. Treebanks, for example, are richly syntactically annotated corpora. The primary source of a treebank is written text, enhanced by a rich syntactic annotation. In acoustic corpora, the primary data source is an audio-recording which is annotated with a transcription and linguistic features. Multi-modal corpora or sign-language corpora are composed of video-recordings as primary data source and their linguistic annotations. Such corpora might be said, according to the point of view of the researcher, to represent a language or a type of languaging.

If, however, a sampling of languaging is aimed at, the use of multimedia corpora becomes inevitable, even for corpus types for which previously pure textual representations have been thought to suffice. The pure transcription of a sign-board for example, can only represent part of the languaging act. One might, of course try to come closer to the event by a complex annotation of colors, forms, orientations etc, using a photo or a video as reference against which annotations are optimized. These multimedia however are best also included in the corpus to support future annotations, to validate the existing annotations and to conduct qualitative analysis on yet unannotated information. We thus foresee, that in the future, newspaper corpora are based on images of an article, which shows, among others the position on the page, associated images and tables as well as colors and fonts. The transcriptions, that are the basis of more traditional corpora will, of course, still be available, maybe linked back to the place of each word or character in the image.

As the use of multi-media allows accessing and analyzing a great variety of social phenomena that go beyond linguistic production, multi-media corpora allow enhancing disciplines with new perspectives on their research: Quantitative analyses become possible where previously qualitative analyses prevailed. Geo-referenced and temporal data allow for the precise delimitation of phenomena and, as a consequence,
to control the generalization. In addition, through temporal and geo-referential data, different but related corpora can be superimposed to handle hence intractable problems. If a public corpus is properly constructed, thus, it can be used for many other research questions than those the developer thought of, even across different disciplines.

In this article, we describe how the notion of public corpus can be extended to the research field of linguistic anthropology. We will first define the term linguistic anthropology as understood in this article, and then, develop a research question from which a corpus-building project is derived. Then we will introduce and illustrate the various levels that we think are generally necessary for the construction of a corpus, i.e. media, indices, transcriptions, descriptions, annotations, references and reference types. Then we will discuss corpora in Linguistic Landscapes and show how our design of a corpus matches or meaningfully extends to this and potentially other corpora.

2. Research Background

2.1 Linguistic Anthropology

Linguistic anthropology is closely related to fields such as semiotics, linguistics, anthropology, anthropological linguistics, sociology, sociolinguistics, and, as we argue, corpus linguistics. Possible borderlines between these disciplines run along the qualitative/quantitative border, separating e.g. anthropology on the qualitative side from sociology and sociolinguistics on the quantitative side. The object of research can be related to language (linguistics, anthropological linguistics), or culture (anthropology, linguistic anthropology). More precisely, linguistic anthropology can be defined as the study of semiotic data, including language, to understand culture-related questions of social identity, socialization or conceptualizations.

Recent analyses of cultural identity have highlighted the socially negotiated basis of an identity (e.g.  


Building a Corpus 8

Brown 2004), as well as the attempts to create of influence social identities through ideologies. Such ideologies come disguised as innocuous semiotic material, such as texts, images, maps, legends, opera, museums, films etc. (e.g. Brown 2004, Teng 2004, Guy 2005, Muyard 2008). Social identity forms thus a typical research question in linguistic anthropology: How are people brought to think who they are through semiotic and especially linguistic exchanges. The interest in research on ideologies comes from its Janus-like nature: On the one hand identities are needed to glue a society together and to prevent ultimately societies from falling apart or civil war. On the other hand, people are constantly manipulated in their social identities, frequently according to the political and economical interests of the influential classes or their counter movements. For Taiwan, these questions are of particular interest.

2.1. Linguistic Anthropology

Taiwan, where our research started, is a multilingual and multi-ethnic society which lives a constant quest of its identity. Having gone through times of Austronesian cultures, Dutch an Spanish colonizations, spreading Chinese Holo and Hakka cultures, Chinese Qing Dynasty colonization, the Japanese colonization, a Chinese Nationalist dictatorship, different periods of Christianizing, a Taiwanization movement from the 1980th onwards, a increasing internationalization of its citizens, and finally a reappraisal of Chinese roots after the presidential elections in 2008, the quest of identities defies any simple answer and is strongly bound to ethnicity, region and age. Two main ideologies compete for the Han-Chinese citizens of Taiwan. The Kuomintang ideology stresses the Chinese origin of the current Han-culture in Taiwan. The opposite, Taiwanese Localization Movement, stresses the fact that the people of Taiwan are descendants of male Han-Chinese and female Austronesian ancestors and that this amalgam created a distinct Taiwanese culture. Both ideologies are written from a Han-perspective and don't question the basis of this Han-ethnicity. Both ideologies don't recognize the fact that the 'Han-Chinese' that moved from the south of China to Taiwan before 1895, were themselves the offspring of, in the mother-line, aboriginal groups, that were sinicized in the South of China by rape and marriage, similar to the sinicization of Austronesian aborigines in Taiwan (Xue et al. 2008). Thus, both ideologies stick to
the Han idea that "it was patrilineal kinship that survived beyond death; matrilinel ties (through one's mother) and affinal links (through marriage) were generally terminated through death" (Watson 1988, pg. 8). The first, Kuomintang, ideology is a coherent, but a quite selective Han ideology. The second ideology, is more considering, but criticizes the Han ideology from a Han point of view, i.e. accepting a pure Han status of the Chinese Southerners, but refusing a pure 'Han' status for the Taiwanese.

The current Taiwanese society has seen accordingly, throughout the basically unbroken Kuomintang administration, the establishment of an omnipresent machinery that pushes each individual into a Han-identity, to the extend that foreign children with blond hair and white skin declare themselves 'Chinese' while attending a Taiwanese Kindergarten, so the German 5-year old son of a German non Chinese speaking couple working in Taiwan (personal communication). Media, schooling, Chinese script, face, guanxi, a traditional calendar, traditional Chinese holidays, Fengshui, Buddhism, Taoism and Confucianism are the cornerstones of a society that structures each individual life from the cradle to the grave from a Han-perspective.

### 2.2. Claims of Identity on Taiwanese Tombstones

Even after death, the Han-identity is marked on the tombs and the tombstone, not only through Han-specific symbols and gods and the Chinese script, but through a specific field on the top of the tombstone that marks the 'affiliation' of the deceased. These affiliation fields, however, show a variability, which is barley known in and outside Taiwan. The different types of affiliation that we observed are.

* A reference to an ethnicity or dynasty, found till now in China only, e.g. 'Han'
* a very detailed address in China, found in Indonesia and with Mainlanders in Taiwan,
* a name of a region in China, mostly of 2 or 4 characters,
* a Tanghao,
* a place name in Taiwan, mostly of villages or city districts,
* a historical place name in China,
* a marker of a religion,
* a reference to a military division,
* a generation number, or
* an empty field.

As the content of these affiliation fields is determined consciously and, to some extent, aware of the variation of affiliation of types and tokens, we must think of these fields are claims that are expressed in view of the reader implied by the text, either uttered by the deceased or by his or her social environment. Remarkable is that again, non-standard Han-Chinese identities, e.g. Christianity or Taiwanesity, are claimed through a standard way Han-Chinese claim their identity. But how can we understand the identities claimed? Can the Taiwanese identity be said to be in contradiction to a Chinese identity? How are the identity claims supported or contradicted by other components of a tombstone and the tomb, e.g. symbols, images, architecture, gods, writings, offerings, orientation etc. How do these claims vary over regions, ethnicity, time periods and socio-economic backgrounds? These questions can be tackled only through a corpus of tombs and tombstones, including geo-spatial and temporal data as well as linguistic and non-linguistic signs.

Theoretically, we could have built a private corpus, selecting some tombs island-wide in a random way and marking in a notebook what is written on the top of the tombstone together with some symbols we find on the tombstone or the tomb. However, it was obvious from the beginning that we would fail to keep track of all necessary information at the first visit, so that we would have to revisit island-wide the tombstones anew, or start after each insight the research anew. Actually, after one year of field trips in Taiwan, we always came back with new insights about tombs and relevant pieces of information that, we think, have to be included in a related research. With a simple private corpus we would, after one year, still be at the very beginning, starting over and over again. In addition, building a private corpus represents, on a more global scale, a waste of time, money and energy. Searching graveyards and accessing tombs is a time and labor-intensive work, requiring long journeys, suffering from sun, rain,
wind and mosquitoes. Finally, this work is not without risks such posed by landslides, swamplands, insects and snakes. To minimize efforts and risks, not every researcher with a related research question should have to start a similar island-wide field-trip. A public corpus is thus mandatory. It allows us first to go back at any stage in our analysis and consider more and different data, if this information has been captured in the photos. Second, interested researchers should be in state to tackle research questions, linguistic or anthropological, related to tombstones and tombs, even if they have not been involved in the corpus-building project. How such a public corpus is to be build is the topic of this paper.

3. Elements of a Public Corpus

The question we primarily tackle is thus, how a public corpus that features a segment of a culture, should be constructed, so that the data can be used easily by a wide range of researches for potentially many research questions. We will tackle this questions first using examples from our tombstone corpus and than try to show the generality of our in comparison to other corpus types.

3.1. Media Files

The primary data of a corpus should be as close to the reality under investigation as possible. The primary data thus should be multi-media representations of the reality. We suggest thus, that even standard linguistic corpora should start from images of texts, e.g. scanned pages in order to be able to fall back to another level of description whenever necessary. Having as primary data only a transcription or a description would filter out too much of what later might be needed by researchers.

Of course, taking pictures or videos does not guarantee loss-less capturing of data. Some data cannot be captured, e.g. what is behind a stone, under the earth or the color under the color. To minimize the loss of information, however, the procedures by which data are collected in the field are very important. We will come back to this procedure. But even with a sub-optimal procedure of taking images or video or a sloppy field work, a lot information is collected unintentionally and can be used later on. For example, using
Building a Corpus 12

color photos, we can go back at any stage of our research and consider the color of this or that element for our analysis without paying too much attention on that while taking photos. Like the color, a lot of other features can be taken unintentionally, e.g. the relative size of elements, the material, the orientation of figurative works, the shape of the tombstone and many more. If the primary data would be a transcription or description, no data would be collected unintentionally.

Our corpus is thus composed mainly of photos. With our corpus type, videos offer no advantage for the documentation of tombs as they are not moving. In addition, photos are easier to handle than videos. The disadvantage is that photos represent the object in different fragments that have to be linked together in a structure that represents the object under investigation. With less clearly structured objects that change rapidly in time, videos are of course mandatory, e.g. for rituals or linguistic exchanges. Our corpus includes some videos and audio recording of people explaining tombs and graveyards.

The procedures by which photos are taken are guided by the following ideas: We do not reconstruct the graveyard, i.e. we ignore the interaction between tombs, paths and temples, as we think these relations not to be central. If it would be easy to include them, however, we would have done so. The relation between the tomb and its natural environment, e.g. mountains, rivers is certainly of significance, but difficult to capture except through satellite photos. The positions of tombs in terms of WGS 84 data (NIMA 2000), however, are known through a GPS and are integrated into the corpus. Tombs thus can thus be merge with a Geographic Information Systems (cf. Streiter 2007, 2008) or tools like Google Maps or Google Earth. If our GPS-device would allow recording the orientation of the tomb, e.g. North or East, we would have collected that information as well and reconstruct the relation between the tomb and rivers and mountains in a GIS. Our GPS devices however do not capture the orientation together with the locations.

The position of elements of a tomb however should be represented in the photos, so that for each object we basically take 2 photos, one showing its position in its environment, one showing its close-up if interesting. If objects are manipulated for photo-shooting, e.g. putting together a broken tombstone to make it readable, the first photo should represent the original status, e.g. the scattered tombstone, the
second the reassembled tombstone. In principle, this would apply also for the removal of plants on or around the tomb which hinder the view. As this would result, however, in thousands of simply green photos, we don't follow this directive rigorously. As a consequence, the status of the tomb such as 'overgrown' or 'abandoned' (cf. DeBartolo Carmack 2002) cannot always be reliably reported.

One global front photo and on global side-photo should localize most elements of the tomb. In some cases one single photo from one corner of the tomb suffices to localize most elements. Elements not visible or not predictable, or too small in the general view are then taken in detail. The order of this photo-taking should help to reconstruct the tomb from the individual photos. Generally speaking, we first take photos of the center, than proceed from left to right. The point of view of the photos is basically that of the mourner, e.g. a person facing the tomb. The orientation of the deceased, imagined as sitting with his or her back leaning against the tombstone, or that of any god or goddess would be possible and meaningful as well. However as the point of view of the mourner can be applied most systematically and is the most informative, we mainly assume this point of view.

There are different techniques to get the GPS-date into the multimedia. One is to use a GPS-device and take a photo of the GPS-device together with the object. The low readability of the photographed GPS-device and a lot of hand-coding limit this approach to small corpora, where potentially extra photos of the GPS-device are taken. A better way is to synchronize a GPS and the camera and store the path taken on the field trip. Then, for each time of a photo the GPS data can be calculated from the path (cf. Erle, Gibson and Walsh 2005). Still better is a GPS enabled camera that stores the GPS-date in the TIFF or JPEG header of the image file. One can also work with a GPS-enabled camera and one or more non GPS-enabled cameras on the same object. If the cameras are synchronized, the GPS-date can be inherited from the GPS-enabled camera. GPS-enabled audio- or video recorders, to our knowledge have not yet been developed, but files could inherit their GPS-date from a synchronized GPS device or a synchronized GPS-camera. Actually, we do not synchronize two devices literally, but measure their time difference. The time difference is than used to calculate the GSP-data for the not GPS-enabled device.
In addition to the photos, audio-files of interviews, as well as video-files documenting the field-work itself can be added to the primary data. All these media-files have to be connected in a form to reconstruct the object under study.

3.2. Indices to Media Files

As a prerequisite for annotations and investigative queries prior to systematic annotations, the media-files should be searchable, similar to a Google image search. With more complex media files, the search should allow to refer to sub-parts of a media-file. E.g. the lower part of an image, or a video from second 2 to 6. We thus need an index which consists of two parts, a searchable term and media-file names and their sub-parts, for example, "broken tombstone" --> 11556.jpg,58484.jpg,38474.jpg,294857.mp3.

Indices are technically easy to realize, are quick in annotation, e.g. need just a yes/no answer and should allow a rough classification of the media files with respect to their content. Indices, thus differently from the annotation, do not refer to a specific tomb or tombstone but firstly and and in an unstructured way to phenomena identified on the photo. Using indices, photos can be easily handled during all processing steps. Indices, for example, should prepare and facilitate the annotation of a next layer, or allow for searches in the corpus beyond the annotated structure, without showing thousands of non-relevant images. Indices might thus be created for relatively rare phenomena, for phenomena identified but not yet fully understood so that they can be translated into an annotation or relatively abstract concepts that refer to less than the half of the images, so that by a combination of the index terms, the set of images can be reduced. For example, using the index 'urns' and 'writing' one identifies from the more than 30.000 photos some 50 photos which show an urn onto which something has been written.

The index complements the annotation, e.g. an XML-description of graveyards, tombs and tombstones and thus should not copy the information of the annotation. For technical requirements, the annotation can always be converted into an index atomically. E.g. when the XML-element 'tombstone' links to an image, the image can be automatically indexed with the search term 'tombstone', to make all tombstones
searchable, or when a tombstone has in its transcription a specific word, this word can be converted into an index. We will not talk about these automatically created indices any more.

Indices currently used are the following: For abstract concepts to delimit the search space we use: Writing, figurative work, abstract ornament, abstract symbol, offering, photo, deity, and decorative ceramics. Indices pointing to specific phenomena not yet understood or not yet annotated are: tombstone_broken, tombstone_manipulated, urns, coal, tudigong_broken. Indices with the pure technical function to speed up the annotation process are gps-device (the GPS-device is on the photo), Tudigong, Longshen (the respective goods are on the photo).

3.3. Structures and Inverted Indices

The structure is a digital representation of the objects, their parts and relations in parallel to the media-files. Usually an XML-structure is developed and used, TEI (CITE) and XCES (CITE) for texts, and TSML (Streiter, Voltmer and Goudin 2007) for tombs and tombstones. This XML-structure, which can be searched through an input to statistical programs links back to the image-files which gave rise to their formalization. Thus, the XML describing a tombstone links back to all images showing the tombstone. It is this XML-structure into which most annotation labor is invested and that slowly by slowly comes closer to the reality as it can be apprehended through the media-files. This basic structure of objects and their sub-part is then completed with transcriptions, descriptions, references and reference types.

3.4. Transcriptions

Transcription refers to the insertion of textual elements that are visible on the objects into the structure. Once textual elements copied, they can be automatically searched, parsed and new information can be derived from them automatically. The collected transcriptions of such a corpus can be used as a specialized linguistic corpus, e.g. a corpus of tombstone inscriptions. The transcriptions should closely
reproduce the text in its character color, the writing direction, new lines, whether a character is relatively large or small, set high or low, moved to the right or left etc, so that coarse-grained analyses of the texts can be conducted without the consultation of media-files.

![ThakBong, Keywords in Context](image)

**Figure 1:** Keywords in Context of a multimedia tombstone corpus, showing hits of the character ‘母’. Such a Keyword in Context facility can be used either for research or the management of the corpus, e.g. revealing annotation mistakes.

### 3.5. Descriptions

Figurative work or photos cannot be transcribed but have to be described, in order to become searchable and automatically analyzable. If the figurative work shows recurrent patterns, as it is the case with the Chinese symbols used on the tombs, the description can be tackled with a limited vocabulary (e.g. lion, lantern, dragon, phoenix, lotus). A more detailed description through XML would, of course be preferable, but highly complex. If needed for an analysis it can be added later on. The descriptions than can be converted into indices. Transcriptions and descriptions can be nested, e.g. for a text on a photo used in a text.

Descriptions apply also to shapes, e.g. of the tomb-stone or the space covered by the tomb, to colors, the state of preservation of an object or the material. Most of the time the description is done through a
limited set of adjectives, e.g. *round, yellow, broken* and nouns, e.g. *wood or lion*. Descriptions, differently from transcriptions, presuppose a classification, either a classification of physical properties, the relevance for an analysis has still to be established, e.g. when classifying a tombstone as *marble, concrete or wood*, or well-established classification, e.g. those of Chinese symbols. Sometimes the encoding of descriptions depend on the perception of the annotator, whether he or she perceives something as still 'straight' or already 'bent', as 'gold' or 'yellow'. Such ambiguities might be taken as a starting point for further investigations, e.g. whether or not yellow and gold actually represent the same, idealized color, or whether the angular form of the tombstone is or is not relevant for those who made the tombstone, as they otherwise might have been more clear about the form. Thus, perceptual difficulties in descriptions might indicate that the descriptions do not match the underlying concepts.

3.6. References

References identify those objects that are referred to in descriptions or transcriptions. The reference of a cipher is a number, the reference of a calendar unit is a date, the reference of a place name is the respective place, the name of a deceased reference a person etc. Through the inclusion of the references in a corpus, the objects under study can be included into a network of relations that go beyond the object. It thus becomes possible to relate a tomb to dates, places and people, each of them accessible through other resources, such as encyclopedias, Geographic Information Systems or Chronicles. Whether these references are taken to be the anchored in reality or in another discourse to which the discourse represented by the tomb is linked, may depend on the view of the researcher.
3.7. Reference Type

Reference types represent a classification of the techniques used to achieve references. The distinction of reference type and reference we make here is very close to the Fregian distinction of Sinn and Bedeutung (Frege 1892/1962). While reference and Bedeutung are almost equivalent, our reference type refers to a type of Sinn. Such types can be, for examples, different calendars, number systems, languages, scripts, gender markers or honorifics represent different techniques to achieve the same reference. Reference types are closely related to human conceptualizations and sometimes identities, c.f. Ingulsrud & Allen (2003). Examples of reference types on tombstones have been given in a previous publication (Streiter, Voltmer & Goudin 2007). Here some examples:
3.8. Measurements

Measurement might have been included in the corpus but for technical reasons have not. Retaining the absolute size of tombstone and other parts of the tomb would have interfered with our scope. Additional material and additional field researchers would have been required and thus making the whole endeavor impossible. Size can be important however, e.g. when comparing different tombstones at one tomb, different names on one tombstone, or when comparing, for example Tudigong in Taiwan from the 5 to 15
times smaller Tudigong we found in Hong Kong. Measurements differ from descriptions as the former use continuous scales while the latter use categorical scales. Thus measurements do not directly map onto cultural categorical concepts but might hint to relative concepts, such as importance, hierarchies, explicitness and consciousness. In addition, measurements might tell us also about factual constraints and possibilities, or the time constraints under which an object has been realized.

Figure 5a and 5b: Just a question of size? Tudigong in Taiwan (left) and Hong Kong (right), i.e. the small structure to the left of the tomb.

4. Transfer of Annotation Levels to Corpora of Linguistic Landscapes

Are the different types of annotations we introduced above in view of a tombstone corpus also applicable to other multimedia corpora in linguistic anthropology that one might want to develop? Taking the example of Linguistic Landscapes, as one related corpus type, we will show how the annotation levels would be used in such corpora and potentially enrich the annotation and analysis.

Linguistic Landscapes try to establish the use of languages in public signs, comparing, for example the use of different languages within a multilingual city (Ben-Rafael et al. 2004, Calvet 1990), or its relative development through time, or the more private or official nature of writings in different languages.
Corpora for linguistic landscaping are mostly private corpora, e.g. Backhaus (2007).

The media-files in a public corpus of Linguistic Landscapes, if to be constructed, would consist of geo-referenced photos of streets, houses, walls and signs. These objects form the basic levels for the representational structure, e.g. `<street name='...'> <house type='restaurant'> <sign type=''> </sign> </house> </street>`. The media files have to be classified and indexed, e.g. to input them into different annotation procedures, such as for global street views, for street signs, for advertisements or graffiti. Or, photos have to be classified according to the languages visible on the photos so that photos can be directed to different annotators.

The transcription of the signs would be as obligatory as with the tombstones, allowing for example an analysis of what is written in which language. After the transcription many annotation layers can be added automatically or semi-automatically. A description of images, symbols, forms of boards etc. seems also useful, for example to identify connotations with the choice of a language or the domain of the sign etc. In addition, with Linguistic Landscapes the classification of signs as top-down (governmental) or bottom up (made by citizens) seems to be firmly established as a valuable distinction.

References would refer to the street by a street sign, as a street sign pointing to 'Street X' can be in 'Street Y'. In addition, through parts of the name references to cities or famous people are established, that probably contain political connotations. The *Via Roma* in Bolzano/Bozen shows where the political power is. The Henanlu (河南路) in Taibei and Taizhong, but not in Tainan reflect different traditions of identities, which perfectly match differences in identities as we find on our tombstones.

Reference types of street signs might describe how people are identified, either through their military position, e.g. *Calle Del General Elorza*, Oviedo, Asturias, Spain or by its contribution to literature, e.g. *Calle del Poeta Daniel Florido*, Santa Olalla, Andalucia, Spain. How are references to restaurants made, through the name of the staple food, e.g. 飯/麺, through the name of the food container, e.g. 當, through the name of a restaurant chain or the kind or service, e.g. 動餐. Special days can be remembered in street
names through the date, e.g. *Calle Doce the Octubre*, Cordoba, Spain, or the name of the event, e.g. Christmass Street, Northam, Australia.

Size of street signs and streets is of course an important measure. The Karl-Marx-Allee in East Berlin, for example is with 100 X 2000 meters more presentable than the Hegelstrasse in Magdeburg with 36 x 300 meters.

To sum up, we thus see that the different annotation layers we developed and proposed for the annotation of tombs and tombstones transfer perfectly, for example, to the development of public corpora of Linguistic Landscapes. In addition, this systematic application reveals new research perspectives in Linguistic Landscapes that have till now not been tapped, especially through the inclusion of references and references types into these analyses. Unfortunately, as most corpora are private, a later annotation is unlikely to happen.

5. Summary and Conclusions

Corpora in Linguistic Anthropology can do what Linguistic or Anthropological Corpora still have to fight for, i.e. they can give up any preexisting uniformity assumptions and sample one phenomenon and reconstruct this phenomenon in a temporally and spatially referenced corpus. Through such corpora, which apparently seem to dismantle the notions of 'language' and 'culture', if superimposed through their common references, languages and cultures can be reconstructed, although in an empirically validated and less ideologically biased way. It is thus perfectly conceivable to superimpose or even completely merge a tombstone corpus and a corpus of Linguistic Landscapes and to validate conclusions against each other.

In this paper we proposed some common layers for such corpora that help with the processing of the multimedia data, and trigger a machinery of linguistic and anthropological analyses. Especially the introduction of the references allows for rich analyses. On the one hand the references and their social or political connotations can form the starting point for an analysis; on the other hand the relation between the references and the linguistic or semiotic means to achieve these references can be analyzed. Why, for
example, in the first author’s German hometown, street names referring to cities far away where noun-noun structures, e.g. *Braunschweig Weg*, while street names referring to nearby villages and cities where adjective-noun structures, e.g. *Lippstädter Strasse*? An analysis of the reference types in relation to the references and the signs yields also rich analyses.

Through these analyses, the references, which have a seemingly encyclopedic or historical nature can be demystified and shown to be used for ideological purposes. Migration, as addressed in this conference and which we capture as a relation between two references, is thus not primarily a historical fact but a discourse that creates ideologies and identities. From all the movements of people, only those charged with ideologies are conceived as migrations.

**Bibliography**


Oliver Streiter, National University of Kaohsiung, ostreiter@nuk.edu.tw
Yoann Goudin, Ecole des Hautes Études en Sciences Sociales, Paris
Ann Mei-fang Lin, National University of Kaohsiung
建立語言人類學語料庫

以Thakbong為例

Oliver Streiter
國立高雄大學
Yoann Goudin
Ecole des Hautes Etudes en Sciences Sociales, Paris
林美芳
國立高雄大學

摘要
語料庫一詞在人類學與語言學領域中，指的是研究資料庫並可用於解答研究問題，譬如說研究某種語言的過去式，那可能就建立一個過去式句子的資料庫；研究特定民族的習俗時，就建立一個照片、影片亦或是訪談筆記的資料庫；通常適用於個人的研究，這類型的資料庫我們稱之為私人資料庫。另一種類型的語料庫，見於語料庫語言學及電腦語言學，這是一個系統化的資料庫，建立此種資料庫的目的為更精確的將資料原意完整呈現，我們稱之為公共資料庫。在此篇論文中，我們將論述如何將此研究方法與人類語言學的領域結合，首先必須先有一個具體的語料庫建立計畫，再者，了解這樣一個語料庫可以涵蓋多少的研究題目與領域。所以如何定義人類語言學也將在本文中說明：針對語料庫討論到的範圍包括資料蒐集的媒介、索引、資料抄寫描寫、資料的註解以及參考資料的建立，再討論如何衡量資料的代表性與意義，這些都將在本文中一一提及。

關鍵字：語料庫、語言學、人類學、人類語言學、語料庫語言學、電腦語言學、自我認同