Empowerment through language description, empowering language description

(Part-II)

Christian M.I.M. Matthiessen

3. Description: sources of data: primary and secondary

The system SOURCES OF DATA is concerned with the data on which the description of a particular language is based, viz. either primary sources or secondary sources, as shown in Figure 1 above; this is the distinction between linguistic phenomena (primary sources) and the ***study*** of linguistic phenomena (secondary sources) — e.g. between "grammar" and "grammatics", as explained by Halliday (e.g. 1996).

> **Primary** sources are instances of the language under description, either authentic texts in the contexts of situation or elicited examples from the language (or speaker observations about the language). The description of the system instantiated by texts is induced from these primary sources (either manually or automatically).
> **Secondary** sources are descriptions of the language under description (ranging from field notes to comprehensive reference grammars) or descriptions of other languages that can provide guidance in the development of the systemic functional description of the language under description. The development of the systemic functional description may involve "translation" of categories in non-SFL description into SFL (Halliday 1996).

The two sources are shown in Table 3 in relation to the distinction in DIRECTION between analysis and synthesis (Figure 1).

Table 1: Intersection of order of sources and direction of investigation

Direction:	Primary sources	Secondary sources
analysis	e.g. manual analysis of clause-rank patterns; automated analysis of morphological patterns	systematic review of existing descriptions
synthesis	e.g. constructing examples based on emerging description to check with language consultant	—

I will discuss the use of primary sources first, and then turn to secondary ones. When we begin to describe any particular language in systemic functional terms, we rely on primary sources: all languages we engage with will be manifested to use as texts in context, i.e. as primary data. In contrast, in the case of many languages, we will not actually have secondary sources, i.e. existing description for all languages; and for those with existing descriptions, the quality and the coverage of the descriptions vary considerably, and we need to interpret them in systemic functional terms.

3.1. Primary sources

Primary sources are either 'authentic' or 'elicited', as shown in Figure 1. I will begin by characterizing the nature of 'authentic' sources, which typically means a corpus of texts; and then I will return to the activity of eliciting primary sources in Section 3.1.3. below.

If we plan to develop a comprehensive systemic functional description of the lexicogrammar of a particular language, we will (ultimately) need a very large corpus, a "reference corpus". If the language under description is a "major" one (e.g. one covered in Comrie's 2018, selection of over 50 languages as "the world's major languages"), it is more likely that linguists have already compiled corpora and developed fairly comprehensive (usually non-systemic functional) descriptions, recent ones increasingly often being corpus-based; the likelihood of large accessible corpora and comprehensive reference description increases according to the financial support provided for a given language. The composition of such reference corpora will, of course, need to be examined systemic functionally; for example, it will be helpful to profile them registerially in terms of their registerial composition and also to determine whether they are based on complete texts (in their contexts) so that we can examine lexicogrammatical features in reference to their "discourse use"[14].

---

[14] The early corpora of English from the 1960s, like the pioneering Brown Corpus, consisted of passages of texts no longer than two to five thousand words, so their semantic-contextual integrity was not ensured; it was in fact

3.1.1. Registerial composition of corpus supporting description

If no corpus is readily available for the language under description, we will need to compile one. I will return to considerations involved in the compilation of corpora below. However, whether a large reference corpus is already available or we need to compile one, it is important to consider the way into the description of the language and the **registerial composition** of the corpus, whether one is already available or not.

Table 2: Examples of registers that can be used in exploratory investigations of certain grammatical systems

Grammatical systems	**Examples of registers in early investigation**	**Characteristics of registers**
experiential: TRANSITIVITY	folk tale / folk taxonomy (taxonomic report) / sequential explanation / verbal map (topographic report) / recipe or other procedural text	variation in the experience being construed
interpersonal: MOOD	casual conversation: dialogue with short turns / teacher-learning dialogue / recipe or other procedural text	variation in the speech roles adopted by interactants
textual: THEME	folk tale / folk taxonomy (taxonomic report) / comparative report / historical recount / verbal map (topographic report) / recipe or other procedural text / exposition (argumentative text)	variation in the method of development organizing the texts

The initial way into the description can or should be based on a selection of simple texts that are likely to expose some fundamental properties of the lexicogrammar of the language under description. I

---

often compromised. They were based on samples of written English, characterized primarily in terms of the contextual parameter of mode, and arguably favouring texts that were easy to sample. Thus there was no guarantee that they were registerially balanced. Their size was one million words, which is nowhere near enough for the development of a reasonably comprehensive description. This was naturally due to technological constraints, and current reference corpora are now tens of millions or even hundreds of millions of words in size. If a corpus of a given language is still composed of texts that are limited to a few thousand words, with longer texts being represented only in part, it is advisable to collect complete texts in their contexts representing different registers so that it is possible to explore, identify and describe grammatical systems "organically"; this certainly applies to textual systems such as THEME and REFERENCE (cf. Rose 2001b).

give some examples in Table 2 for each metafunctional clause system likely to be relevant to the language under description.

So, for example, at an early stage in the investigation of a language under description, we should study dialogues with short turns concerned with the exchange of information and also with the exchange of goods-&-services. In the selection of dialogues, we should thus try to get a good spread of complementary speech roles, and control for tenor variables, e.g. selecting dialogues where the interactants are of equal status in terms of power (status), other things being equal. If we can access such texts, dialogues in books for children and dialogues in textbooks can be a good starting point because they are likely to be simple in a fairly natural way.

In Martin, Quiroz & Figueredo (2021), systemic functional linguists present descriptions of eight interpersonal grammars of eight different languages, using a variety of registers: Monty Python sketch (translated into Spanish), the "Argument Clinic", service encounter dialogue (Spanish, Quiroz); conversation among family members, conversation among colleagues in their office, conversation between government officials and peasants (Zhang, Khorchin Mongolian); courtroom discourse (Wang, Mandarin); family discussion among siblings of wedding plans, dialogue between dressmaker and client (Martin, Tagalog); spontaneous dialogue taken from corpus of Brazilian Portuguese (Figueredo); casual dialogue between two signers familiar with one another, elicited data collection (Rudge, British Sign Language); dramatic dialogue from a novel (Bartlett, Scottish Gaelic). As contributions to the book show, dialogues from a range of different registers, either 'authentic' or 'elicited', can serve as a basis for developing and illustrating the description of the interpersonal grammars of particular languages. I would recommend specifying the values of the contextual parameters of field, tenor and mode for each of the registers (functional varieties of language) represented in the sample of texts upon which the description is based; see e.g. Zhang (2020).

If we sample dialogues in Arabic with short turns and high interaction, we would begin to be able to discern aspect of the clause grammar of system of MOOD (and also the verbal group grammar of MODE). For instance, we would find evidence for a systemic mood distinction in 'free' clauses between 'indicative' clauses and

'imperative' ones, with the mode of the verb depending on whether the POLARITY of the clause is 'positive' or 'negative' (with different markers of 'negative' polarity depending on clausal mood but also on process type). Within 'indicative' clauses, we would find a distinction between 'declarative' and 'interrogative', and we would further find a distinction within 'interrogative' clauses between 'polar' interrogatives, marked by *hal* or *ʔa-* at the beginning of the clause, and 'elemental' interrogatives', characterized by a Q-element placed at the beginning of the clause: من *ma*, ما *ma:*, ماذا *ma:ðа*, كيف *kayfa*, أين *ʔayn*, متى *mata:*, لماذا *lima:ðа* and so on. The placement of these interrogative markers would be a possible indication that the beginning of the clause in Arabic is a marker of thematic prominence, an indication that would be taken up separately in the study of texts with very different "methods of development" (cf. Fries, 1981), like narratives vs. taxonomic reports (the former being more likely to be "VSO" and the latter more likely to be "SVO").

3.1.2. Registerial composition of the corpus, and windows on different grammatical systems

Elaborating on the theme of the registerial composition of the corpus that is used to support the development of the description of a particular language, I will take one more step, and introduce a **field-based map of context** with observations about the **meanings and wordings "at risk"** (e.g. Halliday 1978). This field-based map represents ***just one projection of the map of context***; in the pursuit of language description, it must be complemented by tenor-based and mode-based maps in order to foreground interpersonal and textual resources, respectively, "at risk". For example, using a tenor-based map in developing a description of a language with lexicogrammaticalized distinctions relevant to the enactment of differences in power ("vertical relations", or "status") and familiarity ("horizontal relations", or "solidarity") in e.g. Korean, Japanese, Javanese, we can gradually flesh out systems of "speech level" and their interaction with systems of MOOD (e.g. S.E. Martin 1992; Sohn 1999; Teruya 2007; Kim et al. 2023).

The field-based map is set out in Figure 6. The areas of the grammar "at risk" in the different regions of the map are merely examples, and obviously there is considerable variation across different languages "embedded" in their unique contexts of culture.

Thus, the map is only one source of ideas relevant to our decisions as to how to proceed when we select texts relevant to the description of distinct regions of the content plane of a particular language. The general methodological point is that text from different registers can serve as "gateways" into specific regions of the content plane — into the meanings and wordings that are "at risk".

For example, assume that we are planning to develop a description of the temporal grammar of a particular language by trying to determine whether it is modelled in terms of (1) TENSE, or (2) ASPECT or (3) a combination of these two models of the processes unfolding through time or (4) no such systems (each of which is characteristic of different regions of the world, as we can see by searching WALS or Grambank). When descriptive and typological linguists began to work systematically on tense-aspect systems, drawing on evidence from discourse in the 1970s, their register of choice was, naturally enough, the traditional story: a register operating in contexts characterized by the field of activity of recreating: narrating. This is quite useful, because we can then ask how languages with different temporal lexicogrammatical resources enable speakers to tell stories (cf. Longacre 1974, 1990; Rose 2005, 2024) — see e.g. Bohnemeyer's (2009) account of temporal anaphora in a "tenseless" language, Yucatan Maya.

However, to get a more well-rounded picture, we need to examine texts instantiating registers other than traditional narratives[15], registers where the flow of time is a central concern. As indicated in

[15] In his survey of tense and aspect systems, Dahl (1985, 198-2006) used a different method — a questionnaire of "sentences" and "connected texts" with some indication of "context". The "connected texts" are actually only short passages from narratives, e.g. "(B1) [Do you know what happened to me yesterday?] (161) I WALK in the forest. (162) Suddenly I STEP on a snake. (163) It BITE me in the leg. (164) I TAKE a stone and THROW at the snake. (165) It DIE." The "sentences" in the questionnaire are numbered (1) through (197). Dahl (1985, 37) describes the questionnaire as follows: "The main part of the questionnaire consists of a number of sentences and short connected texts in English together with indications of the contexts the sentences or texts are assumed to be uttered in. These sentences and texts were then translated into the languages to be investigated by native informants."

Figure 6, these include registers that operate in contexts that can be characterized in terms of the following fields of activity:

- recreating: narrating — (traditional) stories
- sharing: reminiscing — stories of personal experience
- reporting: chronicling — (historical) recounts / real-time recording of events (e.g. sports commentary) / forecasts (e.g. oracular texts, weather forecasts)
- expounding: explaining — sequential explanations
- enabling: instructing — procedures

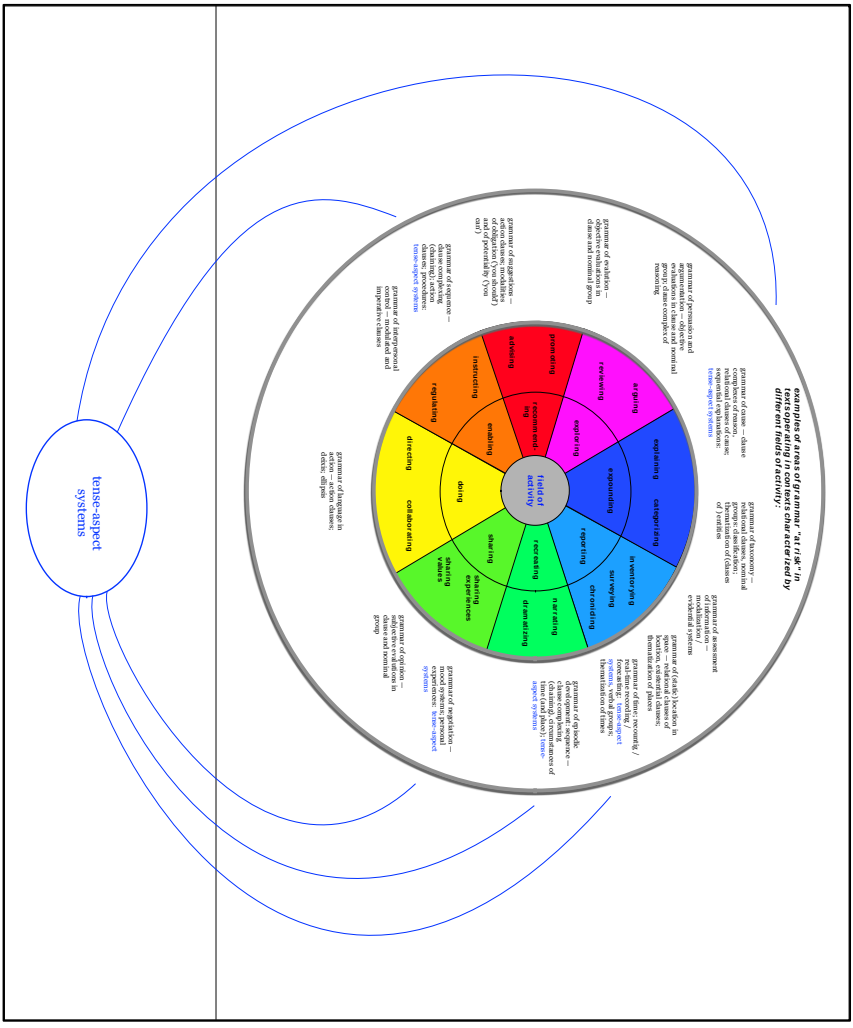
This list is certainly not exhaustive, and we must allow for the possibility of fields of activity of special interest in particular contexts of culture; but the registers ensure that we cover different time frames (eras, periods, episodes, events), the contrast (if relevant) between factual (e.g. recounts) and imaginary (e.g. stories), the temporal ranges from past via present to future (as temporal notions, not as grammatical categories; e.g. both recounts and forecasts), the distinction between instantial (as in stories and recounts) and generalized (as in sequential explanations and procedures).

By moving in "from above", from the characterization of context (more specifically, field of activity), we can discover ***variation across languages in the registerial solutions they have evolved***. For example, in English, recipes are constructed as sequences of instructions realized by 'imperative' clauses, but in Arabic they are constructed as sequences of unbounded culinary operations realized by 'imperfective' clauses in the 'passive' voice (without agents specified, the systemic default in the language). Giving registers a central place in the overall accounts reflects the insights into register variation across languages explored and reported by e.g. Teich (1999), Lavid (2000), Murcia-Bielsa (2000), Lavid, Arús & Moratón (2009), for example demonstrating differences in the interpersonal area having to do with tenor, speech function and mood[16]; see also Matthiessen (2023, 184-186).

---

[16] This strand of investigation may also be compared with work outside SFL in cross-cultural pragmatics on "speech acts" (e.g. House & Kádár 2021, Chapter 8).

Figure 6: Fields of activity and examples of associated registers likely to put certain wordings at risk and thus provide textual gateways into the study of those areas of grammar



For example, in investigating the temporal grammar of MSA, we can compare and contrast texts from three registers operating in contexts characterized by different field of activity, as shown in Table 3. In sequential explanations and in recipes (as an example of a procedure)' the 'imperfective' aspect is unmarked, but in (traditional) stories, the 'perfective' is unmarked. Importantly, we then need to go on to examine the use of the more marked aspectual

option; this is perhaps most obvious in stories: the 'imperfective' tends to be used in the representation of simultaneous events that are not bounded in time (but which are not "located" in the present). This division of labour between the two temporal options strongly suggests that the basic system is one of ASPECT, not one of TENSE.

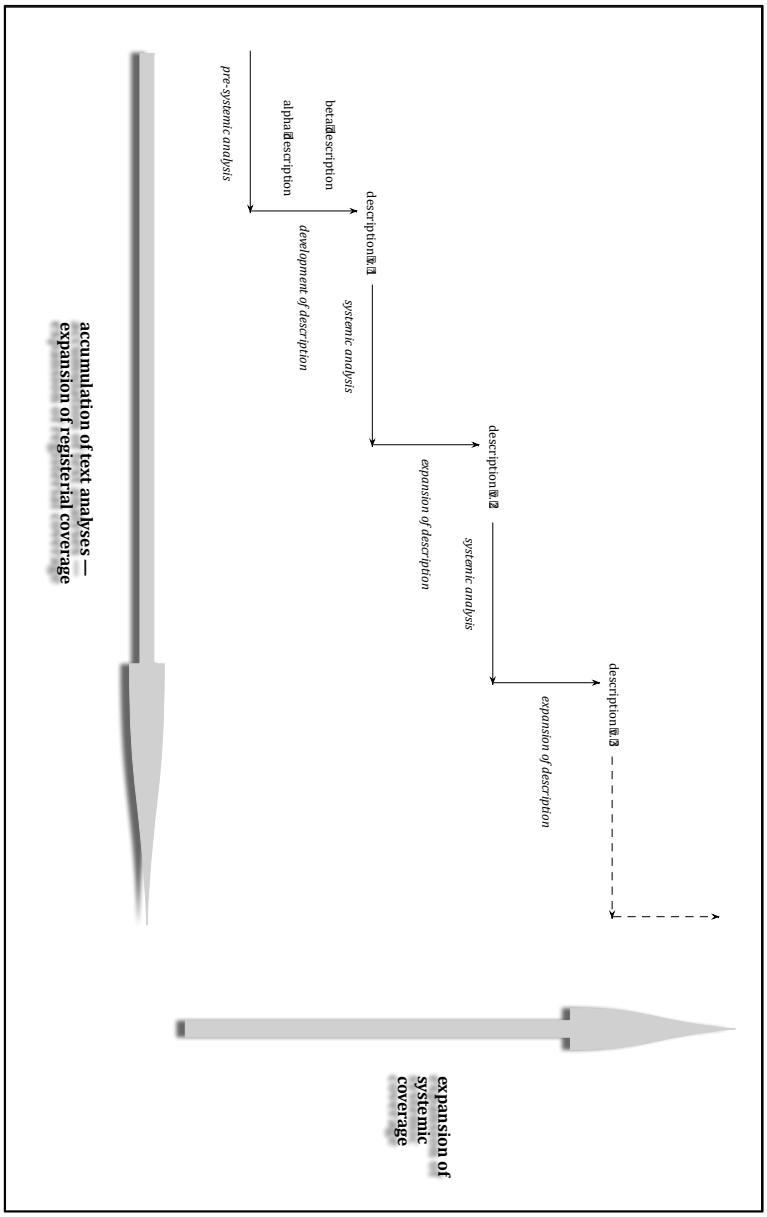
Table 3: Examples of registers in the investigation of the temporal grammar of MSA in relation to unmarked and marked aspectual choices

Context: field of activity	**language: register (functional variety of language)**	**unmarked aspect**	**other aspectual options**	**comments**
expounding: explaining	sequential explanation	imperfective		
recreating: narrative	(traditional) story	perfective	imperfective	main event line: perfective; simultaneous events, not bounded in time: imperfective
enabling: instructing	procedure, recipe	imperfective		clauses are passive in voice, without specification of the 'agent' (which is characteristic of MSA)

3.1.3. Stages in the development of systemic functional descriptions

Once we begin to sketch the description of (some facet of) the grammar of a language, we can test it against the analysis of additional text. In this way, the development of the description of a particular language can be characterized as a succession of descriptive versions, each being an expansion of the previous version with corrections based on testing against the analysis of additional texts, as shown in Figure 7.

Figure 7: Stages in the development of systemic functional descriptions



The first stage of analysis is ‘pre-systemic’ (see Figure 1) since we have not yet begun to construct the description of the system of the language in focus (i.e. the system located at the potential pole of the cline of instantiation) — unless of course we have been able to

interpret a non-SFL description of the language in provisional systemic functional terms (see Section 3.2 immediately below), and can use it as a first draft of a description — what I have called an "alpha description" in Figure 7: we can think of the successive versions of the description we are developing much like the successive versions of software — an important aspect of the testing being the use of the expanding description in the analysis of additional primary data, i.e. of new texts.

Working towards the first draft, we may find it helpful to use computational tools that enable us to detect lexicogrammatical patterns "from below". For example, given the tendency for the most frequent lexicogrammatical items to be grammatical items ("function words") rather than lexical items ("content words"), such tools can provide us with a preliminary list of likely grammatical items. Thus, according to AntConc, the most frequent items in Cervantes' Don Quijote are *que*, *de*, *y*, *a*, *la*, *el*, *en*, *se*, *no*, *los*, *con*, *por*, *le*, *lo*; these are all grammatical items. The first lexical items are the title *don* (rank 23), the names *Quijote* (29) and *Sancho* (30) and then the verb *dijo* (35) followed by *respondí* (50). This thus gives us an **item-based index of the grammar**[17]: the most frequent items are determiners ("articles"), (structural) conjunctions (linkers and binders), prepositions and pronouns. (I have found that if I am working on a language that I do not know as a speaker, identifying and then trying to learn grammatical items can be very helpful; for example, it will be useful in producing interlinear glossing.)

As we expand the systemic coverage at the potential pole of the cline of instantiation and the coverage of texts from additional registers, we may both increase the delicacy of the description and also add simultaneous systems. For example, systemic functional linguists often start with a textual thematic description of the clause as a way into the lexicogrammar, since it often (but not always) also provides a "window" on aspects of the interpersonal and experiential clause grammars. Once we have reached a certain degree of descriptive delicacy, we may then decide to begin to explore simultaneous systems like MOOD and TRANSITIVITY. However, the

[17] This can serve as a "dictionary view" of the grammar; cf. Matthiessen (1995) for different views on the lexicogrammar of a language.

easiest points of access into the lexicogrammatical system vary across languages; and if we have access to secondary sources, to non-SFL descriptions, this will probably be reflected in the balance of their coverage: they are likely to give more space to the more exposed overt parts of the grammar, and to overlook more cryptic covert aspects. And if the language under description has a reasonably elaborated word grammar ("morphology"), we may have to sort out the basic principles early on in the descriptive project; and secondary sources are likely to devote a large part of the description to morphology, as Ryding (2005) does in her reference grammar of MSA.

3.1.4. Elicited (vs. authentic) texts

In the discussion so far of 'primary sources', I have assumed that the texts being analysed are 'authentic' rather than 'elicited': a corpus is a sample of authentic texts — texts occurring naturally in their contexts of situation, selected for inclusion according to explicit criteria. However, as we develop a description of a language, we will almost certainly also need to 'elicit' texts or fragments of texts. We elicit them from "native" speakers of the language — language consultants[18] we are working with; this a common strategy in linguistic fieldwork (e.g. Newman & Ratliff 2001; Vaux, Cooper & Tucker 2007; Bowern 2008; Chelliah & de Reuse 2014; Thieberger, 2014; and for phonetics in particular, Ladefoged 1997). If we are "native" speakers, we can also consult ourselves (although it is still advisable to consult other speakers). This will typically happen when we cannot find naturally occurring examples to support the development of the description of the language we are concerned with (probably because they are quite rare); or we can also use elicitation as a shortcut to produce a provisional description, one which we later test against 'authentic' texts.

---

[18] They were often called "informants"; but this term has negative connotations so although it is sometimes still used, the term "consultant" is preferable — and it indicates the expertise of the speakers helping with the descriptive project (cf. e.g. Vaux, Cooper & Tucker 2007), and it serves as a reminder that language consultants are potential future descriptivists. An excellent example is the well-known collaboration between Franz Boas and George Hunt; see e.g. Wilner (2015).

We may elicit whole texts to supplement our corpus of naturally occurring texts, for example asking our language consultants to produce versions of very familiar texts like texts used in ceremonies or folktales — broadly speaking, texts that are part of the discursive cannon of a community. Alternatively, we can stage a context of situation, and ask our consultants to play their parts; this may be helpful if we want to collect dialogic texts, e.g. casual conversation. The prompt may be fairly detailed, or it may simply be a topic that relates to an important event in the community[19].

When we elicit whole texts, the unit of elicitation is defined in terms of context (settings of the values of the field, tenor and mode contextual variables) since texts are characterized "from above" as language functioning in context (e.g. Halliday & Hasan 1976; Halliday 1981); so that is why I said that we can "stage a context" (cf. House & Kádár 2021, on the methods used in cross-cultural pragmatics, including the pioneering speech act realization framework). However, as we develop lexicogrammatical descriptions, we will also find that we need to use units and unit complexes from the lower stratum of the content plane, i.e. from lexicogrammar. That is, the environment of elicitation is one stratum down within the content plane (for the notion of "environment", cf. Matthiessen 2001).

Thinking about such elicitation based on lexicogrammatical domains (units or unit complexes), we can use it to explore single systems such as TRANSITIVITY, MOOD and THEME.[20] For example,

---

[19] For example, as we guided the compilation of OZTalk, a corpus of spoken Australian English, Di Slade and I invited a group of friends to reminisce about their experiences of a freak hailstorm in Sydney. We recorded their conversation in a soundproof room in the Macquarie Linguistics phonetics laboratory so that we would have top quality recordings. It turned out that the interactants quickly forgot about the artificial material environment, got quite engaged and animated sharing hailstorm stories. The recordings were of high quality, certainly good enough for phonetic analysis to support phonological analysis (cf. Halliday & Greaves 2008).

[20] Here it is important to note that the success of elicitation based on grammatical domains will almost certainly vary across metafunctions. Among the three metafunctions (ideational: logical & experiential, interpersonal, and textual), the textual metafunction is most likely to be negatively affected by elicitation based on grammatical domains. Such elicitation has contributed to

elicitation can be a quick way of developing a sense of how the language under description models (i.e. construes by means of the system of TRANSITIVITY) the experience of possession, attribution of quality or of meteorological phenomena (these being domains where languages vary in how they model the domains). However, such elicitation often involves probing systemic possibilities beyond single systems — i.e. probing paradigms defined by the ***intersection*** of simultaneous systems such as AGENCY and PROCESS TYPE, MOOD and POLARITY, MOOD and PROCESS TYPE.

Such **paradigm probing** can be an excellent way of detecting systemic interdependencies and realizations conditioned by terms in more than one system. For example, by probing the intersection of the systems of MOOD and POLARITY in English, we will find that 'exclamative' declarative clauses are always 'positive' (Matthiessen 1995, 413; 2023a, 241-242) and that the combination of 'yes/no-interrogative' and 'negative' has a special meaning having to do with the speaker's expectation as to the polarity of the answer (Halliday & Matthiessen 2014, 174).

In English, the realization of 'negative' always involves a variant of *not* (unless it is combined with usuality: *never*, or the clause is "quasi" negative: *hardly*; *seldom* etc.). But in many other languages, the realization of 'negative' polarity depends on terms in other systems, e.g. MOOD, ASPECT and PROCESS TYPE, as in Mandarin (e.g. Halliday & McDonald 2004) and MSA, and we can find out precisely what the intersections are by probing paradigms, i.e. by checking intersections of terms from these systems. Thus when we examine the intersections of the systems of MOOD (and FREEDOM), POLARITY and ASPECT in MSA, we find that negative markers are conditioned by combinations of terms from these systems, as shown in Table 12 (discussed in Section 5.3. below). (There is one negative marker not shown in the table, the negative verb ليس *laysa*: it is

---

the seriously misleading myth about "free word order" and "fixed word order" languages (cf. Halliday 1985). There are reports of language consultants producing a succession of examples with different "word orders" in the course of elicitation; but this is hardly surprising if textual influences like "method of development" are not controlled for.

conditioned by terms in the system of PROCESS TYPE; it occurs in 'relational' and 'existential' clauses, as shown in Table 7.)

In terms of trinocular vision, paradigm probing represents the view "from roundabout": we explore systemic combinations both to determine whether they occur or not and if they occur, to discover how they are realized. Obviously, working with 'authentic' texts in a corpus, we can approach this investigation "from below" as well. However, there are a number of reasons why this may be difficult. (i) It is virtually impossible to determine non-existence by only searching the corpus precisely because we can't easily search for non-existent items or patterns. (ii) Certain systemic combinations, like 'negative' and 'yes/no interrogative' in English may be quite rare, so we would need a very large corpus to find a sufficient number of clear examples, and it may be hard to distinguish between non-existence and low probability. (iii) As already illustrated, the patterns realizing systemic combinations, like negative items in MSA, may be varied in word class and form, so we won't necessarily have the full list of items to search for.

By probing paradigms, we can, as noted, explore and determine systemic interdependencies. This includes the basic issue of the relationship between two systems: determining whether they are simultaneous or ordered in delicacy, a relationship which may change over time (as can be investigated in terms of systemic probabilities if we have a diachronic corpus): see Halliday (2004).

The methods of eliciting examples are set out in Figure 1 above. So far, I have touched on variants of what I have called 'field work elicitation', in contrast with 'questioning'. In 'field-work elicitation', we elicit (fragments of) texts, using prompts defined by trinocular vision: "from above" — from contexts or semantics when our focus is on lexicogrammatical description and "from roundabout" — based on lexicogrammatical systems, as in paradigm probing. It's less likely that we will approach the elicitation of lexicogrammatical examples "from below" in terms of the hierarchy of stratification, i.e. from phonology or graphology.

At the same time, we can elicit examples either through 'analysis' or through 'synthesis'. The method that became very prominent in generative linguistics can be characterized as 'synthesis': linguists typically working on their "native" language constructed

examples[21], characterizing them as either acceptable ("grammatical") or as unacceptable ("ungrammatical", presented as starred examples). However, in a typical field-work situation where linguists work on languages they don't know as ("native") speakers (e.g. Bowern 2008), they would most likely ask their language consultants to translate an example from a shared language, e.g. how would you say "it's raining" in Arabic? In other words, language consultants need to use both analysis and synthesis; they need to analyse the example in the shared language, and then synthesize it in their language.

But there are semiotic alternatives to the use of translation in elicitation — notably Kenneth Pike's famous **monolingual**, described in "The Nature of Field Work in a Monolingual Setting"[22] (see also Headland 2004). (I was fortunate to attend on such monolingual by Pike in the 1980s, and both the process and the results were quite impressive. Pike's interaction with the speaker whom he met for the first time on the stage was actually also a useful reminder of how important the tenor parameter of the context of elicitation is — his cheerful engagement, enacting a very positive vibe [this took place in Southern California, where vibes are prevalent and significant].) In the monolingual, Pike would use "body language" and also Mixtec, a language he spoke fluently but

---

[21] As a relevant aside, it is worth noting that the notion of the "native speaker" has been problematized in applied linguistics; see e.g. Davies (2013) and Dewaele, Bak & Ortega (2022). At the same time, "native speakers" are not necessarily good at making grammaticality judgements; one has to practice this skill, learning to imagine possible co-texts and contexts. (In a discussion Michael Halliday and I had about the status of non-finite act clauses as Phenomenon in 'perceptive' mental clauses, he suggested the example *John eating a banana has to be seen to be believed*, which shows that the Phenomenon, the act clause *John eating a banana*, can serve as Subject in the passive version of the clause *(somebody) has to see John eating a banana*. And I remember the great computational linguist Martin Kay pointing out during one of his visits with us at the Information Sciences Institute that there is absolutely no reason to assume that languages have evolved to enable their speakers to produce grammaticality judgements.

[22] URL: https://www.sil.org/about/klp/influence/nature-field-work-monolingual-setting For a demonstration of Pike's monolingual, see also Daniel Everett's presentation at: https://www.youtube.com/watch?v=sYpWp7g7XWU

not one shared with the consultant; this introduced an element of naturalness into the context of situation. In addition, he had a number of objects on a table that he could point to, hold up and manipulate.

In the system network in Figure , the systemic alternative to 'field-work elicitation' is what I have called 'questioning' (about the language): here we would seek information about our consultants' view of the language, e.g. in order to support the development of a sociolinguistic characterization of the language or to gain information about the "folk models" of the language in its speech fellowship (e.g. regarding registers), including "campfire grammars" in oral cultures (cf. Halliday 1977).

3.2. Secondary sources

Returning to the system network in Figure 1, let me now discuss the option of using 'secondary data'. Secondary data are ***descriptions*** of particular languages rather than "raw" data — instances of these languages, i.e. texts in context (primary sources). Such descriptions include ones of the same language as the language under description but also ones of other languages as long as they provide us with information relevant to the development of the description of the language under description. In either case, we can use them as sources of information and guidance when we develop our own description.

3.2.1. Same language

If we are developing a systemic functional description of a language that has already been described by means of other, non-SFL frameworks, it clearly makes excellent sense to try to use them as much and effectively as possible — naturally, depending on the quality of the descriptions; and the descriptive "distance" between existing non-SFL descriptions and the planned SFL description is also a significant factor.

Let's assume that we have access to a reference grammar or another type of comprehensive description when we prepare to develop our systemic functional description[23]. What can we expect

[23] Here I will focus on how to extract information from such descriptions. They will obviously include examples, and often also passages of texts, and we can

to find, and where can we expect to find it in the presentation of the description? In her guidebook to producing descriptions of the grammars of languages, Aikhenvald (2015, 16) recommends the following plan of presentation:

> A grammar starts with an introduction containing basic facts about the language and its social setting, the family it belongs to, and the cultural background. This is followed by a statement of phonology and phonetics. Then comes morphology, then syntax, then sometimes discourse properties and some notes on lexical semantics. Other types of organization are also possible. Syntax may be placed before morphology; it is however important that relevant facts about inflectional morphology be summarized first. Otherwise, the discussion of syntax may become unintelligible. A detailed discussion of phonological processes may appear later in the grammar; however, the phonemes of the language have to be introduced at the beginning, as building blocks for the understanding of what follows. (Aikhenvald 2015, 16)

I have visualized her recommended organization of the description diagrammatically in Figure 8. The sequence of presentation recommended by Aikhenvald is in fact quite common in non-SFL descriptive grammars, so it is also a helpful guide when we search for information that we can draw on when we develop a systemic functional description. Aikhenvald is of course a very experienced descriptive linguist[24]. For example, her pioneering description of Tariana can be used as an example of her descriptive recommendation (Aikhenvald 2003). It contains 23 chapters, starting with "the language and its speakers", moving on to "phonology", and then ascending into grammar, beginning with "word classes" as the

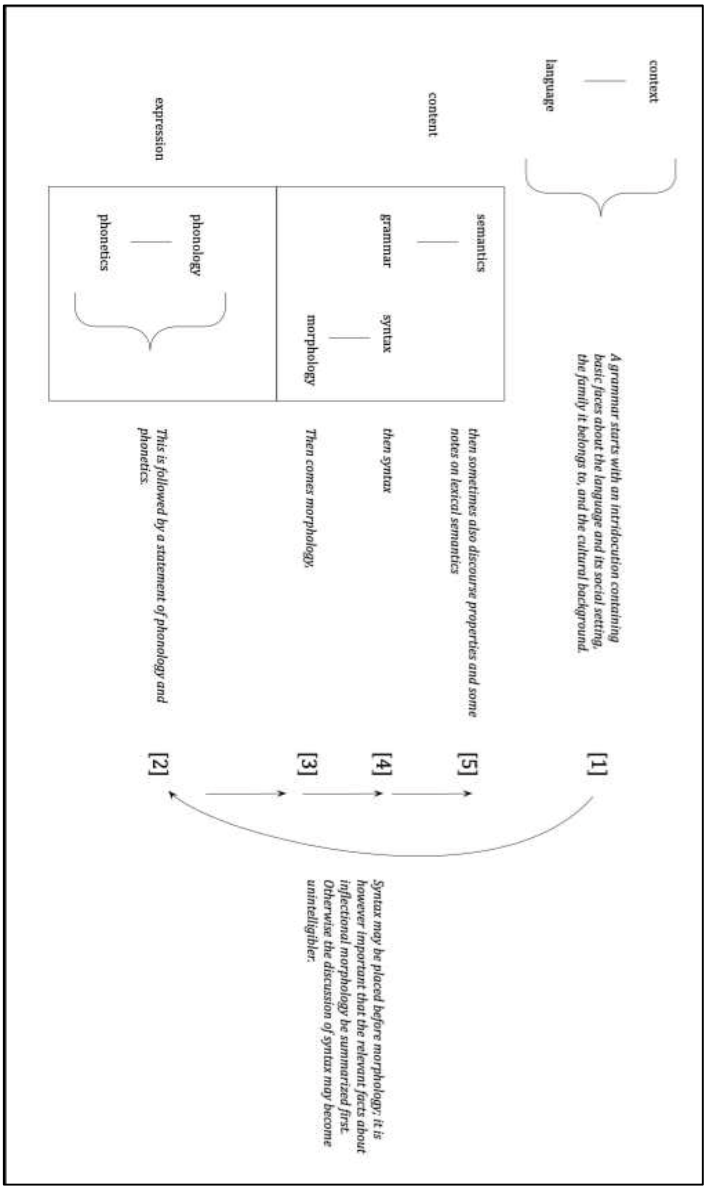
---

treat these as another source of primary sources — included in the secondary sources. In other words, we can use descriptions of languages also to sample texts belonging to different registers, and include them in our compilation of corpora (with proper acknowledgment, of course). Franz Boas and his group included three "deliverables" in their language descriptions: a grammar, a dictionary and texts.

[24] For other non-SFL account of "writing grammars", see e.g. Ameka, Dench & Evans (2006); Payne & Weber (2007); Nakayama & Rice (2014).

first of a number of chapters dealing with morphology, thus following the path of presentation she recommends that is set out in Figure 8.

Figure 8: The organization of the presentation of a description of the grammar of a language according to Aikhenvald (2015, 16), interpreted in terms of the strata and ranks of the systemic functional architecture of language in context



(Source: *Aikhenvald 2015, 16*)

One of the first steps in "harvesting" a descriptive grammar would be to interpret the sequence of presentation along the lines I have done in Figure 8, so that we obtain a clear sense of what we can expect will be covered and what probably will not be. This is likely to give us a rank-based view on the description — with adjacent strata, phonology and perhaps some notes on semantics (including discourse patterns). This represents the vertical dimension of a systemic functional **function-rank matrix**; we will hopefully gain insight into the ranks and primary classes posited in the description, although it may not be rank-based: if it is a traditional grammar, the account may cluster around words and clauses (or "sentences"), with less or no information about groups and phrases (since the model of traditional grammar was one of words in clauses, skipping the intermediate rank of groups and phrases).

The next step is to try discern the different metafunctional regions in the description, attending to the horizontal dimension of the function-rank matrix (Figure 3, Figure 4, and Table 2 above). This is likely to be much harder because the grammar writer will probably not have foregrounded metafunctional distinctions, and when we probe for information about systems within the different metafunctions, this is likely to reveal gaps in the coverage, again because the grammar writer did not work with a sense of goal for the description that included metafunctional considerations (cf. Section 2.1 on Whorf's "plan and conception of arrangement").

Still, it is very helpful to start with an empty function-rank matrix, and try to fill it with references to chapters and sections in the descriptive grammar or grammars we are trying to "translate" into an account that can inform our systemic functional interpretation of the language. As an illustration of this procedure, I have filled in a function-rank matrix for the grammar of Modern Standard Arabic (MSA) with references to chapters and sections in Badawi et al.'s (2016) "comprehensive reference grammar" of written MSA: see Table 6. (For a systemic functional description of MSA, see Bardi 2008; for a description of the experiential clause system of transitivity with a focus on process type in colloquial Egyptian Arabic, see El-Rabbat 1978).

Table 6: Function-rank matrix *of MSA: references to secondary sources*

Rank	**Class**	***General***	**Logical**	**Experiential**	**Interpersonal**	**Textual (1)**	
						Structural	**Cohesive**
Clause	major (free/ bound)	Chapter 3: The basic sentence	taxis & logico-semantic type: Chapter 6: Coordinated sentences; Chapter 7: Subordination Chapter 8: Conditionals	transitivity: Chapter 3: The basic sentence [§ 3.0 Introduction; § 3.1 Equational sentence; § 3.7 Verbal sentence; § 3.11 Transitivity and intransitivity]; § 3.16 ka:na 'be'; § 3.17 Verbs of becoming and duration; § 3.18 Verbs of existence; § 3.19 Verbs of non-existence § 3.20 Verbs denoting 'can', 'be able', 'want'; § 3.20.2 Verbs of wanting, wishing; § 3.29 Dependent noun objects and complements	mood: Chapter 4: Interrogatives, indirect speech (§ 7.9: Verb morphology); § 1.6.10 Interrogatives; § 3.24.1 Positive direct commands; § 3.24.2 Positive indirect commands; § 3.24.5 Prohibitions; § 3.24.6 Indirect prohibitions; § 3.24.7 Coordination of imperatives; § 3.25 Optatives; § 3.27 Exclamatory verbs; § 3.28 Exclamatory sentences with *wa-* … § 3.26.2 Conditional polarity: Chapter 4: Negatives; § 3.6 Negative equational sentences; § 1.6.9 Negatives § 1.6.7 Emphasizers; § 3.30 Restrictives	theme: § 3.3 Topic-comment sentences § 1.6.7 Emphasizers: topic focus (§ 3.3.4)	
group/ phrase	prepositional	§ 2.6 Prepositions					
	nominal	Chapter 2: Noun phrase structure	§ 2.2 Apposition; § 2.19 Coordination	Chapter 6: Adjectival and relative clauses		determination: § 1.5.1 Definiteness markers; § 1.7.2 Demonstratives; § 1.7.3 Relatives	
	verbal		§ 3.21 Verbs denoting 'again', 'still', 'nearly', 'hardly', 'almost'; § 3.22 Verbs of beginning and continuing	tense-&-aspect: § 1.9.3; § 3.10 Aspect and tense	mode: § 1.9.1.5 Mood	voice: § 1.9.1.4	

Rank	**Class**	***General***	**Logical**	**Experiential**	**Interpersonal**	**Textual (1)**	
						Structural	**Cohesive**
Word § 1.4 Morphology	noun	§ 1.8.1 Noun patterns; § 1.11 Noun patterns; § 1.4.6 Pronouns; § 1.7.1 Pronouns; § 1.7.2 Demonstratives; § 1.7.3 Relatives; § 1.8.2 Inflection of nouns; § 1.8.5 Proper names		number & gender: § 1.5.3 Number/ gender markers			
	adjective	§ 1.4.4 Inflection of adjectives; § 1.8.6 Adjectives; § 2.1 Adjectival qualification					
	verb	§ 1.4.7 Verbs; § 1.9 Verb morphology					
	preposition	§ 1.4.5 Nouns as adverbs and prepositions; § 1.6.12 Prepositions; § 1.8.8 Prepositions and prepositionals; § 2.6 Prepositions					
	adverb	§ 1.6.11 Adverbs; § 1.8.7 Adverbials					
	particle	§ 1.6 Uninflected words: particles					
	conjunction (structural)	linkers: § 1.6.3 Coordinating conjunctions (Chapter 6); binders § 1.6.4 Subordinating conjunctions					

(Source: *Badawi et al., 2016*)

My attempt to assign chapters and sections in Badawi et al. (2016) to specific "semiotic addresses" (i.e. intersections of metafunction, rank and primary class) in Table 6 illustrates a number of points relating to the systemic functional interpretation of non-SFL descriptions, some general and some more specific to Arabic:

- Descriptions of systems such as TRANSITIVITY (and within transitivity, the system of PROCESS TYPE) and MOOD TYPE are dispersed into different parts of the overall grammatical description; there is no clear concept of a system defined by its semiotic address in terms of rank (and primary class) and metafunction. For example, the system of MOOD in Arabic (as an interpersonal clause system in the realizational service of the semantic system of SPEECH FUNCTION operating in dialogue in relation to the contextual parameter of tenor) is not described as a unified system. In order to begin to get a picture of the system of MOOD in Arabic, we need to examine certain sections in Chapter 3 — the sections on commands in § 3.24, on optatives in § 3.25 and on exclamations in § 3.27, and also Chapter 4 on interrogative clauses. At the same time, there's also relevant information in § 1.9.15 on the system of MODE in verb (called "mood", as in traditional descriptions in general[25]). But this still leaves us without any clear organic sense of the system of MOOD and the different mood types in Arabic, and its realizational relationship to the interpersonal semantic system of SPEECH FUNCTION as a resource for exchanging commodities (information of goods-&-services) in dialogue.
- The pictures of higher-ranking systems, in particular those of clause and group/phrase rank are thus both fragmented and quite partial in coverage. Just as in the case of the system of MOOD, the system of TRANSITIVITY is only covered in fragments, and the descriptions of it are dispersed. And there

[25] The system of MOOD is the interpersonal clause resource for realizing moves characterized by different semantic speech functions in dialogic exchanges. The system of MODE is also an interpersonal system in MSA but its domain is that of the verbal group, and its systemic terms realize not only clausal moods but also other distinctions, including ones within the domain of 'bound' clauses (cf. Matthiessen 1995, 2004b, 612; Halliday & Matthiessen 2014, 143).

is no systematic connection between the treatment of aspects of TRANSITIVITY in Chapter 3 and the description of the derivational verb forms, or "measures", in § 1.9.

- The textual systems of the clause ("the basic sentence") such as THEME are probably the hardest to get a sense of based on the description presented by Badawi et al. (2016). The system of VOICE can only be accessed "from below", from the point of view of verbal morphology. (Since the grammatical description is concerned with written Arabic, there is naturally no account of the information unit of spoken language.) Any indication of the textual system of COHESION as a unified resource is hard to find, except for § 11.5 on "cohesive iteration". But among linguists describing Arabic informed by SFL, there is work in this area; see e.g. Aziz (1988).
- In general, in terms of Halliday's trinocular vision, the view adopted is "from below"; for example, we can look up different uses of كان *ka:na* as a lexical or auxiliary verb in § 3.16 in Badawi et al. (2016). However, it is virtually impossible to arrive at an understanding of the clause in Arabic as a unification of three metafunctional motifs deriving from semantics — the clause as figure (experiential), the clause as move (interpersonal) and the clause as message (textual). Similarly, in engaging with particular lexicogrammatical regions, we need to view them trinocularly. For example, if we view the temporal grammar of Arabic primarily "from below", as Ryding (2005) does in her reference grammar of MSA, it is difficult to bring out and differentiate the systems of TENSE and ASPECT (thus she interprets the basic distinction in verb forms as tense: present vs. past instead of as aspect: imperfective vs. perfective); but if we follow the approach taken by Holes (2004), we can observe the uses of the temporal categories in their environment in texts belonging to different registers (for discussion of **uses** in the description of a language, see also Section 5.2 below), and also see the significance of the distinction he makes between dynamic and stative clauses (which we can explore in reference to the system of

TRANSITIVITY, with the help of the notion of reactances; cf. Section 3.3).

- While there is a reflection of a description of the nominal group at group rank, there is no systematic separate account of the verbal group (unlike Holes 2004, who has a section on the "verb phrase"[26]); so we have to ***compile out*** information about the verbal group in Arabic based on the description of the verb (morphology) and certain aspects of the clause.
- As noted above, systems at clause rank, and even at group/phrase rank, are not foregrounded; as in traditional grammar, we find paradigms basically only at word rank — conjugational paradigms of verbs, paradigms of pronouns, and so on. By the time the description has moved to clause rank, there are no paradigms defined by the intersection of systemic terms in simultaneous systems, e.g. the systems of POLARITY and MOOD TYPE, POLARITY and ASPECT, and POLARITY and PROCESS TYPE. For example, while the book contains information about different negative items, it is fairly hard to arrive at a clear explicit understanding that the realization of 'negative' polarity in Arabic is conditioned by terms in the systems of MOOD TYPE (including bound clauses) and PROCESS TYPE (cf. the discussion of Table 12 below).

My observations above are only ***examples*** that we can derive from my attempt to view the description by Badawi et al. (2016) through the template ("grid", "lens") of the function-rank matrix — an exercise that I have found very productive as, starting around 1980, I have tried to profile lexicogrammatical descriptions provided by reference grammars of a wide range of languages. Not surprisingly, I always find such descriptions "wanting" in various respects — in large part because they do not embody a sense of comprehensive coverage in terms of system and metafunction, which is of course precisely the view of lexicogrammar brought out in the function-rank matrix. But that does not reduce their value relatively speaking; rather, by profiling the descriptions along the

[26] Although he calls it "verb phrase", it is essentially the verbal group, consisting of verbs and particles; it is not the "verb phrase" in the sense of classical generative grammar (roughly the predicate of traditional grammar).

lines I have just sketched, we can get a good sense of where there is work to be done to fill the gaps and provide an account of the lexicogrammar of a particular language as a resource for creating meanings as wordings. And this will often or always involve moving in "from above", from semantics in context, as demonstrated by Martin (1983) in his investigation of the grammatical resources used to achieve semantic participant identification in unfolding text in three languages, showing that they deploy different sets of lexicogrammatical resources to achieve comparable semantics tasks (participant identification).

3.2.2. Typological databases

In addition to using descriptive grammars as secondary data, we can also turn to typological databases such as WALS and Grambank. They cover a number of parameters (called "features") established for a large number of languages with specifications of the values for each language. In the case of (Modern) Standard Arabic, Grambank includes values for 174 parameters. All of the information has been extracted from Ryding (2005), so the parametric profile of the language depends on her descriptive interpretation of MSA; for example, as noted above, she interprets the basic temporal morphological distinction in the verb in terms of tense rather than aspect. I have set out the parametric profile of Arabic presented in Grambank as Table in the Appendix, adding information about the grammatical domains of the different parameters.

I have characterized the Grambank parameters ("features") specified for Arabic in terms of grammatical domains: clause complex, clause, nominal group, verbal group, adposition. The nominal group and the verbal group are the best represented domains; but actually many or most of the parameters that I have grouped under these headings are concerned with properties at one rank down, i.e. properties of nouns and verbs. This reflects the general orientation in Grambank to the view of lexicogrammar "from below" — rather than "from above" or "from roundabout"; and in the case of Arabic, this is reinforced by their only descriptive, secondary source, which is Ryding (2005). Her "reference grammar" of MSA devotes most space to word grammar. In general, Grambank is necessarily dependent on secondary sources, and the kind of information that is most likely to be documented for a large number

of languages tends overwhelmingly to reflect the view "from below", and Arabic is no exception.

We can use the profile compiled in Table as a source of (distilled) secondary information when we develop a description of MSA. For example, we can begin to discern aspects of the function structure of the nominal group: Deictic ^ Thing (GB025); Numerative ^ Thing or Thing ^ Numerative (GB024); Thing ^ Epithet (GB193). At the same time, we can also find basic information about systems, like DETERMINATION (e.g. GB20-GB23), NUMBER (e.g. GB39-GB44; GB46; GB165-GB166; GB316-GB320), PERSON (e.g. GB28; GB301-GB31) and GENDER (e.g. GB51-GB54; GB192; GB196-198; GB321). However, the greatest value of the profile in Table probably emerges when we develop the description of MSA in reference to other languages relevant to our descriptive project or when we develop a comparative or typological view of MSA.

Methodologically, the approach taken in the development of language profiles in Grambank can be useful to us as we "interrogate" secondary data or even primary data. Each parameter is given with coding instructions for the linguist extracting information from descriptions of the language being coded. For example, for parameter ("feature") GB111 "Are there conjugation classes?" the following specification is provided[27]:

1. Summary
   Are there multiple sets of verbs that each combine with different sets of (inflectional) markers in finite forms? We are concerned with whether there is non-phonological allomorphy of finiteness marking on verbs depending on which verb is chosen. Phonological rules are not considered here but they may correlate with non-phonological allomorphy. Conjugation classes that are lexically assigned are relevant for coding this feature. Please note that a class of suppletive verbs does not count.
   Look for finiteness markers such as: * Person: first, second, third * Number: singular, plural, dual and trial * Gender: masculine, feminine or neuter * Tense: present, past, or future * Aspect: perfect, perfective, progressive * Mood: indicative, subjunctive, imperative, optative * Voice: active, middle, or passive

[27] URL: https://grambank.clld.org/parameters/GB111#2/21.0/151.7

2. Procedure

3. Code 1 if finite verbs are inflected differently depending on what verb is used.
4. Code 1 if there is a two-way distinction between 'regular' and 'irregular' verbs and there are more than two irregular verbs.
5. Code 0 if differences in the forms of TAM, indexing, or voice markers for finite verbs result purely from morphophonological alternations.

We can consider developing comparable questions designed to help linguists extract information needed when they produce a systemic functional description of a particular language[28]. The answers could be recorded in a version of the function-rank matrix being constructed for the language, thus sorted in terms of metafunction and rank. By another step of supporting systemic functional linguists developing descriptions of particular languages, we could build this into a linguist's computational workbench, with information available also from databases such as Grambank.

In addition to systemically compiled and presented typological databases like Grambank, we can of course learn from descriptive generalizations that appear in overviews of language description and typology, including Comrie (1981); Shopen (2007a,b,c); Payne (1997); Whaley (1997); Haspelmath (2001a,b); Moravcsik (2013); Aikhenvald & Dixon (2017); and Dixon's (2010a,b, 2012) "Basic Linguistic Theory", and the many thematic volumes edited by him and Aikhenvald, as well as many other descriptive-typological thematic volumes that have appeared since the 1970s — including overviews of grammaticalization.

The rapidly growing literature in these related areas can certainly inform the development of new systemic functional description. For example, they will include phenomena that have only come to attention and into focus in the last few decades, like evidentiality

[28] Personally, I am reminded of all the work we did as part of the Penman text generation project formulating inquiry operators as part of choosers associated with lexicogrammatical systems in order to make semantically informed choices among the terms of those systems (cf. Matthiessen & Bateman 1991). In the context of language description, such questions would be formulated to discover the information needed about the language to make descriptive decisions (cf. also Longacre 1964).

(Chafe & Nichols 1986; Aikhenvald 2004), mirativity (Delancey 1997), associated motion (e.g. Koch 1984; Guillaume & Koch 2021) — see also Evans (2022) for additional examples. Such accounts are significant because as we set out on developing the description of a new language, we need to have as rich an imagination of what is possible as possible — a reservoir or at least a pool of linguistic possibilities ("etic pool"). Bowern (2008, 214-218) provides a checklist of morphosyntactic topics to explore in fieldwork through elicitation. For further discussion, see Matthiessen (2024).

3.2.3. Other languages (comparison)

Descriptions of languages other than the one under description can also serve as sources of insight. In very general terms, we can use them as descriptive models; they can guide us in the development and organization of the description of a particular language (which is how Halliday's *Introduction to Functional Grammar* has often been used, and sometimes my *Lexicogrammatical Cartography*, Matthiessen 1995), providing alternatives to, or at least variations on, Aikhenvald's (2015) plan, interpreted in Figure 8 above. Importantly, a systemic functional presentation of the description of a language would probably include by way of introduction a comprehensive overview of the language after the introductory chapter, description of the different metafunctional strands within the language, and probably a chapter including systemic functionally analysed texts from a good spread of registers (cf. Halliday 1972, on a "good description").

However, in addition to such insights into how to organize the presentation of a systemic functional description, we can usually find particular descriptive insights directly relevant to the language under description. For example, if the language under description is MSA, we can gain insights by examining descriptions of other particular languages that are (typologically) similar in one respect or another[29]. Thus, drawing on existing systemic functional

[29] Keeping in mind that we base typologies on linguistic systems, not on whole languages — an approach argued for already by Halliday (1959-60). Based on the typological "settings" of systems in a particular language, we can then go on to investigate if they form a syndrome of related features, a basis of a

descriptions of various languages as we develop a description of MSA, we can gain insights from:

- descriptions of German (Steiner & Teich 2004), French (Caffarel 2004, 2006) and Japanese (Teruya 2004, 2007) in order to gain insight into the complementarity between the MOOD system of the clause, as also the system of FREEDOM, and the MODE system of the verbal group (and of the verb) and the complementarity in this area between the ranks of clause, group and word;
- descriptions of English (e.g. Halliday & Matthiessen 2014) and of Mandarin Chinese (Halliday & McDonald 2004) in order to shed light on the interpretation of the grammatical system of processes time in MSA: does it model the unfolding of the process through time as TENSE, as English does, or ASPECT, as Mandarin does, or as a mixture of the two?
- descriptions of Japanese (Teruya 2007), Korean (Kim et al. 2023) and Telugu (Prakasam 2004) and of English since the first three are characterized as "SOV" languages in so-called "word order" typology and English is "SVO" in order to explore the most common sequences we find in MSA (when "S" is present), viz. "VSO" and "SVO". By examining the comprehensive descriptions of the verb final languages, we can begin to get a sense of what to expect when the clause of a language ends with the Predicator/ Process as an interactive event (a dialogic move) — the interpersonal finale of the clause and we can ask if there is a "mirror image" of this in MSA clauses that are verb initial. The answer will be a clear "yes"; the beginning of the clause in MSA can serve as an interpersonal overture to it as a dialogic move. By also considering the description of English (or of another so-called "SVO" language), we can explore the possibility in MSA that clauses may be either "VSO" or "SVO", and we can detect a "split" reflecting distinct "methods of development" characteristic of different registers (e.g. a

---

characterology of the language of the kind Halliday (2014) proposes for Mandarin Chinese.

taxonomic report is likely to be dominated by "SVO", but a traditional narrative by "VSO"). At the same time, we can also gain significant insights from Martin's (2004) description of Tagalog, a language that has been characterized as "VOS" since although the status of the "S" needs to be problematized, Martin's description the verb-initial nature of the language will help us deal with MSA as a verb-initial language (which it is usually characterized as, even though "SVO" also occurs).

When systemic functional descriptions of languages with different typological characteristics are compiled and compared (possibly aided by a profile based on a typological database, as in the Grambank profile of MSA compiled in Table below), it is possible to see how languages have evolved different strategies for "mapping" the range of metafunctional resources onto one another, e.g. with respect to the clause: the clause as a figure, the clause as a move (proposition/ proposal) and the clause as a message. The picture that emerges will bring out **competing motivations**. I tried to show this in Matthiessen (2004b), and now, two decades later, we have access to many more rich descriptions of particular languages and relevant findings from the typological literature.

3.2.4. Liberating oneself from secondary sources

If the language under description is one that has hitherto not been described, or only sketched in field notes, the main challenge for us in embarking on a systemic functional description of the language is to be attentive to the ***primary*** source of data, texts in context, empowered by systemic functional theory and stimulated by systemic functional descriptions of languages that are comparable to the language under description in one way or another. Thus, we can avoid the temptation of imposing the categories of traditional grammar on the languages — one that Tozzer (1921) argues against in his description of the grammar of Maya (in the Boas tradition), reflecting the state of descriptive affairs a century ago.

However, if the language under description has already been covered in non-SFL descriptions — in particular, if there is a long well-established descriptive tradition (as there is in the case of Arabic, starting with Classical Arabic), we may in fact find it harder to "liberate" ourselves from the tradition documented in secondary

sources like reference grammars[30]. I will illustrate aspects of the problem in Section 5, where I discuss briefly the description of a few areas of the grammar of Modern Standard Arabic, noting that the tradition (whether Arab-Islamic or "Western") tends to adopt the view "from below", in terms of axis (syntagmatic) and rank (words [in clauses]). For example, it is essential to transcend the "received" description of MSA clauses as being either "nominal" or "verbal".

There are two key strategies we can use to ensure that we are not as it were trapped in the tradition, but can transcend it:

- **Trinocular vision:** while established descriptive traditions often foreground the view "from below" since this view is focussed on the more easily observable facet of lexicogrammar (explicate order as opposed to implicate order: Matthiessen 2023a, 256-258), we need to adopt trinocular vision, supplementing the view "from below" with the views "from above" and "from roundabout", thus producing a more well-rounded description of the language under investigation.
- **Sources:** the tradition is of course a secondary source, and we can also free ourselves from its take on the language under description by drawing extensively and systematically on primary sources, on texts in context, analysed and described trinocularly — including primary sources like examples and texts provided in the secondary sources, but re-interpreted in the light of SFL. The descriptive agenda should ultimately be determined by texts in contexts, and not by the tradition enshrined in secondary sources (and thus not be issues that have become popular points of debate).

Together, these two strategies should help us avoid exploring the language under description within the frame established by the dominant descriptive tradition. It may even be necessary to set the tradition aside completely for a period of time, and proceed as if the

[30] This includes liberating ourselves from their conception of what should be covered in the description of the lexicogrammar of the language. This conception is likely to be quite limited, again reflecting the view "from below". We need to refocus on the lexicogrammar, viewing it trinocularly as a resource for creating meanings as wordings.

language under description had never been described before; this can actually be a very powerful productive strategy, helping the systemic functional linguist truly perceive patterns in primary sources, i.e. texts in context.

Once we have established a balanced, trinocular view of the language, we can characterize the view adopted in previous descriptions within the tradition, e.g. by locating the components of the description within a function-rank matrix (as in Table 6 above); it will almost certainly turn out to be quite a partial view of the language. As always, adopting trinocular vision means ***shunting*** along the relevant semiotic dimensions to adopt different vantage points, thus avoiding getting stuck with only one perspective on the language (cf. Halliday 1961).

As far as lexicogrammar is concerned, we are not simply producing a systemic functional variant of descriptions in the prevailing tradition. Instead, we actually aim much higher; our goal is to develop a comprehensive, meaning-oriented and text-based description of the lexicogrammar as a resource for creating meanings as wordings — a rich description that can be applied within a range of significant institutional settings outside linguistics, e.g. in education systems, healthcare systems, legislative systems, administrative systems, legal and judicial systems, financial systems, and media organizations.

3.3. Summary; the distance from primary sources to generalized secondary sources

In this section, I have discussed how to use primary sources of data, i.e. texts in context, and secondary sources of data, i.e. existing descriptions, either of the language under descriptions or of other languages, either ones that are similar in one respect or another or generalizations based on samples of descriptions of other languages, including large typological databases. While I have presented these two sources of data as alternatives in the system network of activities by linguists in Figure 1 above, they are of course ***complementary*** — we can re-enter the system network and choose options more than once from the system network, and the degree to which we use one or the other or both will depend both on the state of accounts of the language under description and the context of the descriptive project — the goals, questions and problems we set out to address (the

"consumer" concerns; cf. Halliday 1964). Thus, by re-entering the system network of options in methods, we can characterize so-called "mixed methods research"; but I prefer to characterize them as "complementary" to bring out the fact that they will enable us to investigate and illuminate different aspects of the language under description (cf. Matthiessen & Teruya 2024, Chapter 4).

I have highlighted the systemic functional approach to primary and secondary sources. One central aspect of systemic functional descriptions that I have not brought out is their sensitivity to and reliance on insights derived from Whorf (1956), prominently the use of **reactances** in the development of the description of a particular language and the attempt to reveal **cryptotypes** (e.g. Shore 1996, in relation to Finnish; Rose 2001a, in relation to Western Desert; Wang 2020, in relation to Tibetan; Quiroz 2020, in relation to Spanish). While extensive analysis of 'authentic texts' may produce clues that we can interpret as reactances, we will almost certainly need to engage in **paradigm probing** using some kind of elicitation to detect reactances and gradually reveal cryptotypes. (For example, in the investigation, how do we arrive at the significance of the distinction between *I think so* and *I know* [*it*])? See e.g. Halliday & Hasan (1976) and Halliday & Matthiessen (2006). SFL enables us to locate reactances and cryptotypes within the total system of language. They are located within the 4th layer of grammar, the "inner layer of grammar" discussed by Halliday (1990/ 2003):

> But if we probe into the inner layer of the grammar, to the cryptogrammatic fourth level that people are least aware of, here we find a gradual, clause by clause synthesizing of a world view, a hidden theory of experience on which we unconsciously base our actions and our strategies for survival. There is a syndrome of grammatical features which conspire — in Martin's term — to construe reality in a certain way; and it is a way that is no longer good for our health as a species. Let me try and identify four of these, beginning with one that is familiar and easy to access. (Halliday 1990/ 2003)

This is part of the **implicate order of** language in contrast with the **explicate order**. Here I am using a differentiation of kinds of order taken from Bohm (1980) in his account of order in physical systems (for the full systemic functional interpretation in relation to language, see e.g. Matthiessen 2023a, 256-258). Implicate order is

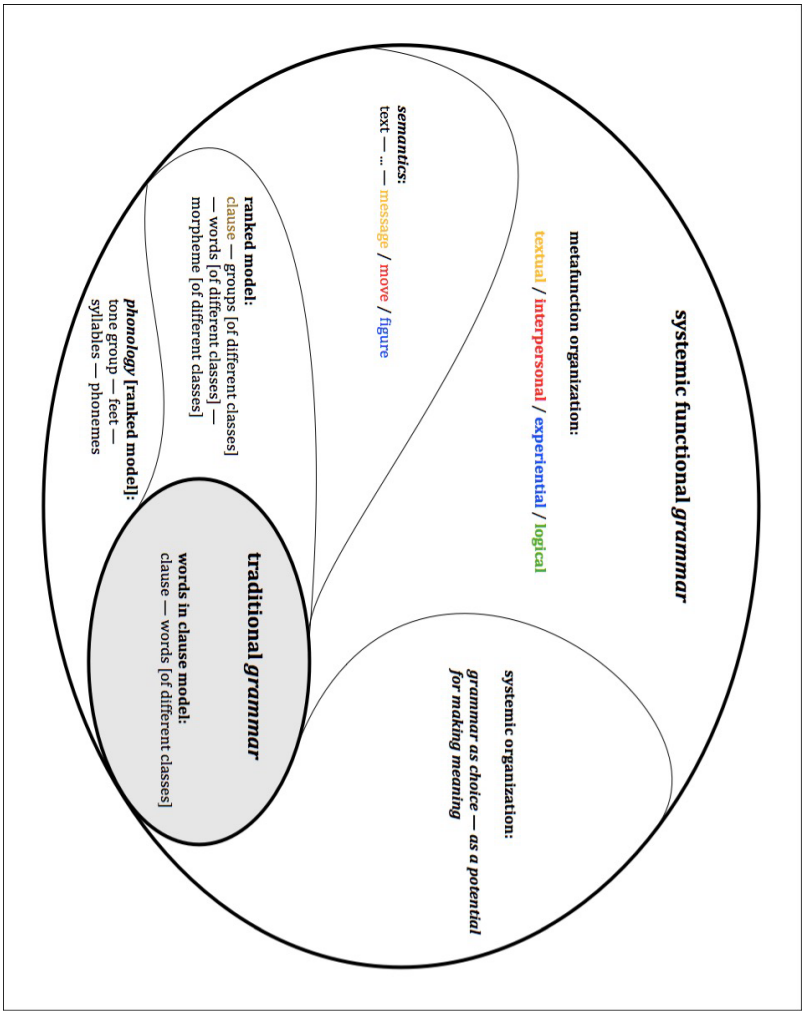
systemic order — patterns along the paradigmatic axis, whereas explicate order is structural order — patterns along the syntagmatic axis. Syntagmatic patterns are more exposed and thus easier to observe than paradigmatic ones, so they are more likely to be covered in descriptive grammars of various languages. But it is implicate order that is generative — in engenders explicate order. So cryptogrammar is embodied in implicate order, as are the reactances that we use to explore and describe cryptogrammar (see also Section 5.1 below).

In short, compared with traditional theories of grammar, and their more current versions in the descriptivist tradition, systemic functional theory of lexicogrammar enables us to observe and describe much more of the overall resources of lexicogrammar as a resource for creating meanings and wordings: see Figure 9.

As shown in Figure 9, the foundation of traditional grammar (in the sense of traditional theory of grammar based on the Graeco-Roman tradition[31]) was the model of grammar as words in clauses (or "sentences"), and the approach was "from below" — having identified different forms of words, traditional grammarians to ask how they were used, and what they signified (cf. Halliday 1977). In contrast, in SFL, the ***theory*** of grammar — or more accurately, of lexicogrammar — is part of a ***holistic theory*** of language in context: the location of lexicogrammar within the overall "architecture" of language is quite clear and explicit, as is the internal organization of lexicogrammar as a stratal subsystem of language, as the lower stratum of the content plane of language. Thus, in SFL, the organization of lexicogrammar in terms of axis and rank comes into view, as does its natural relationship to semantics, embodying the metafunctional spectrum of modes of meaning.

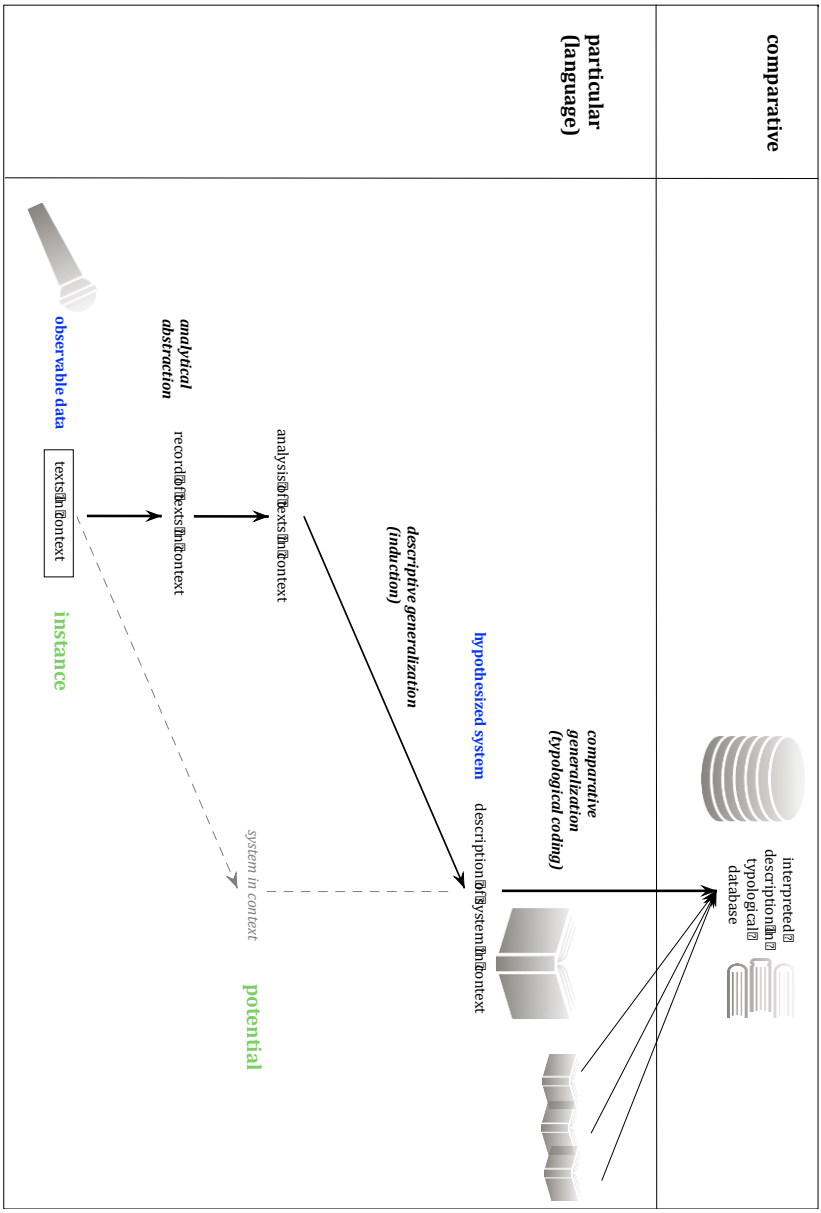
[31] With the important note that in this tradition, there was no clear distinction between theory and description. As a result, many missionary linguists looked for and imposed the descriptive categories of languages they were familiar with, like Latin, on other languages they encountered around the world.

Figure 9: The overall territory of the lexicogrammar of a language that is illuminated by traditional grammar vs. systemic functional grammar



Rounding off my discussion in this section, let me say something about the distance between primary data — i.e. observable data in the form of texts in context — and descriptions based on texts in contexts, and comparative generalizations. I have represented this "distance" schematically in Figure 10.

Figure 10: The distance from data (texts in context in a particular language) to generalized descriptions in typological database



I have shown the distance between our primary data, the texts in context that we can observe, and our generalizations based on the data, descriptions of the systems that lie behind texts in context, in terms of two dimensions that derive from our systemic functional theory of language, viz. the cline of instantiation and the representation of texts in context:

- the cline of instantiation extends from the instance pole to the potential pole. At the instance pole, we can observe our primary data, texts in context; and we can use some form of representation to record these texts, e.g. video-recording, transcription in real time. (Decisions made in transcription are obviously informed by theory; cf. Halliday 2002.) We analyse texts in context, and based on our analyses, we move upwards along the cline of instantiation towards the potential pole, ***inducing descriptive generalizations***. The outcome of this process is a description of the language — our hypothesis about the system of the language. This move involves generalization of the inductive kind; and in the process, we may need to set aside instances that at a certain stage of development of the description appear to be "outliers" for some reason or another (here quantitative information will be helpful). The distance we need to cover is considerable, of course; and as we develop the description of the system, we should increasingly be able to predict instantial patterns, and test such predictions by means of the analysis of additional texts (cf. Figure 7 above, which represents successive versions of the description under development).
- at the same time, there is also a considerable distance between the texts that we can observe, and our successive ***representations of these texts*** as we move to the form of representation that we can analyse (either manually or automatically). Even if we focus on the lowest stratum of the expression plane (phonetics or graphetics, or sign), our representations, e.g. narrow or broad transcription, and analytical tools will inevitably provide a selective grid or "lens"; and if we focus on the content plane, lexicogrammar and semantics, we need to consider carefully how the text is represented to support the kind of analysis we want to undertake. For example, for the purposes of analysing texts

to describe content-plane systems, we may use standard orthography, standard orthography plus prosodic annotation (e.g. intonation and rhythm), or standard orthography plus grammatical or semantic annotation[32]. But all these representations constitute abstractions of information from the texts that we can observe. So we need to determine how to "preserve" relevant information (instead of idealizing it out of the record) but at the same time avoid cluttering the representation of the texts with information that is not relevant to the task at hand and is thus likely to distract us. The general problem of representation of texts as data for analysis and description is of course familiar in linguistics, and is reflected in well-established distinctions such as broad vs. narrow transcription, emic vs. etic representation. As always, our choices here will need to consider the goals of our project (echoing Halliday 1964, I would characterize the choices as a matter of "transcription and the consumer")[33];

---

[32] In corpus linguistics, researchers have articulated different assessments of the value of annotations, either as enhancing the value of text as data or as skewing the representation of the text as raw data. But on the one hand, even the representation of text within the expression plane is a theoretical decision (see Halliday 2002); and on the other, when text has been annotated, we are never forced to include the annotations — we can set them aside if warranted. — If we add annotations, it is important that they should be separable ("stand-off") from the "raw" representation of the text. For example, the very valuable pioneering corpus of spoken British English presented by Svartvik & Quirk (1980) is very hard to use in its original published form because annotations were not separated from the "raw" transcript, so the transcript is very cluttered with information (cf. Halliday 2002).

[33] Thus the original transcription of conversational texts in Conversation Analysis (CA; Sacks, Schegloff & Jefferson 1974) contained both too much information and too little information for the development of descriptions of English lexicogrammar and semantics (cf. Halliday & Plum 1985; Halliday 2002). (1) CA transcripts were designed to represent conversation as interactive behaviour rather than as the exchange of meanings, so micro-behavioural patterns were carefully recorded — patterns which might distract from the analysis of conversational texts as processes of meaning realized as processes of wording. (2) At the same time, CA transcripts did not contain sufficient information about prosodic patterns, so vital information about intonation and rhythm was simply lost in the transcripts produced of conversations.

> for example, if we are working on a description that is intended to bring out the division of labour between language and gesture in the making of meaning, we may need to include information about the timing of units and prosody that we would not otherwise include in a "purely" linguistic descriptive project to enable us to add timing link in the representation of spoken gesture and gesture (which is possible with an annotation tool such as ELAN[34]).

In terms of the two dimensions just characterized, there is thus a considerable distance between observable data, i.e. texts in contexts, and descriptions based on the observable data. So we need to recognize that any descriptive categories that we postulate as we develop in the course of a particular language will be subject to this "distance": they constitute generalizations and abstractions from patterns that may be detected in our observable data, texts in context.

When we take the additional step of comparing descriptions of particular languages, and then, based on such comparisons, we make comparative or typological generalizations, we increase the distance even further. This was something Firth and Halliday knew based on the multilingual engagement in the Firth-Halliday tradition, and they drew attention to this as both a methodological and theoretical issue (e.g. Halliday 1959-60; Halliday, McIntosh & Strevens 1964; Ellis 1987)[35].

---

[34] https://archive.mpi.nl/tla/elan

[35] Their insights are echoed implicitly by Haspelmath (2010, 2019). He argues that descriptive categories posited for particular languages and "comparative concepts" need to be distinguished. I say "implicitly" because he does not refer to them (cf. Matthiessen et al., submitted).