Tito ORLANDI

Towards a Computational Grammar of Sahidic Coptic.

In this paper I wish to present the rationale and the first results of a study, the goal of which is not immediately clear, it may even seem futile, therefore it needs some preliminary words of justification. First of all, I would recall the importance that the *corpora* of texts in electronic format are gaining both in view of the dissemination and preservation of literary texts, as opposed to the printed form, and of the linguistic study of them. Nothing better that the remarks of two great scholars in text linguistics, J. McH. Sinclair, and R. de Beaugrande, may illustrate such point. They are found in two articles, which would deserve lenghty consideration, but the ideas that are interesting for us here may be synthesized in a couple of sentences. Sinclair maintains that “many of well protected assumptions of language analysis are suspect and probably due for radical revision” (p. 4). Beaugrande observes that “the antipathy to data *is* a reflex of the widespread aspirations in theoretical linguistics to replace real language with ideal language,” and that “these revisions can finally be guided by reconnecting real language to the real texts made accessible by very large *corpora* of authentic text and discourse” (p. 244).

But for a *corpus* to be really instrumental to the linguistic research, it cannot be consituted only by rough text. After years of intensive research in this field, it has been established that *corpora* are to be arranged so that each word is accompanied by its grammatical classification, especially including the indication of the part of speech to which it belongs. Otherwise it is not possible to build some software for the analysis of texts, which give satisfactory or interesting results, e.g. an authomatic analysis of Coptic texts, which classifies the sentences according to their structure, and classifies the internal sub-structures of one sentence, or the search for more advanced linguistic analysis, such as the filling of valences, the use of certain structures for certain words, etc..

It is the famous problem of the POS tagging (part of speech tagging), which cannot be conceivably done by hand, and for which a number of automatic parser has been created for the modern languages (notably English, German, and French) but never for Coptic, nor, for that matter, for other ancient languages.

The creation of a Coptic *corpus*, of a Coptic POS tagger, and of a parser for the analysis of the Coptic language, is the general idea behind the computational grammar which I present here. It is in fact necessary to formulate a very precise formal grammar in order to design the software for textual analysis. A computational grammar might be defined as the effort to explain the grammatical structures of one language to the computer (S. Emmel). This gives a fairly good idea of the issues at stake, mainly that we cannot refer to some general experience or intelligence (in the sense of extra-logical intuition) of the reader, and that we must not jump any passage; but it is not a sufficiently formalized definition, which certainly is required here.

First of all, with regard to the adjective "computational," we accept the analysis of computation done by A. Turing, and reformulated by H. Rogers in the shape of the theory of recursive functions (see bibliography), and we maintain that computation is the use of logical procedures (or operations) on formalized data. The formalization of data consists in the identification of defined units, and their mapping to arbitrary symbols, which are used for computation. From this point of view, a computational grammar, or parser, may be defined as *a machine which analyses textual units (sentences, utterances), by means of a vocabulary and of a set of formal rules, and gives as an output the structure of the sentence.*

As to the content of the grammar, being the Sahidic(-Coptic) a dead language, its features can be recognized only from the analysis of texts which we find written in existing documents. That analysis is based on two axioms (otherwise justified by philosophical theories in phenomenology and semiotic, which we do not intend to discuss here), two couples of contradictory principles (justified by the fact that every language is itself an ambiguous phenomenon), and two additional principles (commonly agreed in modern linguistics).

Axioms

a) The utterances of a language (Saussure: *parole*) are composed of units (basic units), in turn composed of units. The analysis of the language identifies those units at the different level of structure (or better, at the different organization of system, in the sense of the General Systems Theory), and the grammar describes the units and the rules which govern their assemblage.

b) The basic units cannot be identified but by means of *contrast* with the other units (contrastivity).

Contradictory Principles

a 1) The text cannot be wrong

a 2) The text may be wrong.

Principle (a1) means that the "text", as we have it, is the ultimate source of the knowledge of the grammatical rules and of the verbal elements (morphs) which constitute the Coptic language. If one sentence contradicts our assumptions, either we must change those assumptions, or we must admit that they do not cover all possible sentences.

Principle (a2) means that the "text" from which we derive our assumptions is not exactly what we find in the document, but the abstract text in the mind of its author, which in the course of being written or copied or dictated may have incurred in incidental mistakes. The range of such mistakes goes from inadvertently dropping a word to inaccurate spellings, etc.

b1) Only the strictly formal aspect of sentences and morphs is taken into consideration, without any assumption about the underlying meaning (context free).

b2) The sentences are considered as the means to convey a message, generally consisting of the predicate of something for some subject in some circumstances.

The formal construction of words and sentences is dependent on the abstract form of the message, in the sense that they are meant to fill in the parts of the message. On the other hand, the rules governing the formal construction of words and sentences may be considered as a closed, independent system, which does not need any extraneous justification.

According to principle (b2), in order to formulate a message, it is required some kind of predicative construction, formed by the subject and by what is predicated of it, and eventually by contructions filling in the valences of the predicate, and by some kind of circumstantial, adverbial construction.

Additional Principles

a) The language is considered in its *synchronic, vs. diachronic* manifestation. No analysis is supported by historical considerations.

b) The grammar is *descriptive, vs. prescriptive*. This depends above all on principle (a1).

Grammatical Structure of the Sahidic Coptic

Textual units

The structure of the Coptic sentences is tendentially consistent: the verbal structure (including the subject) is in first position, eventually followed by the object, then by the circumstantial determinations or complements. The syntax is shaped according to the Greek mental attitude, because Coptic was born in a very hellenized environment, and initially Coptic texts were translations from Greek. Most of the conjunctions come from Greek. Anyhow, Coptic syntax is still to be studied extensively, also in relation with Coptic style, the study of which has never been done.

The sentences may be classified as utterances containing a predicative pattern. They follow one another without superpositions of elements, and are composed of elements called **clause, phrase, word, morpheme**.

The patterns and the phrases are build with "terminal" categories of words, and also possibly with other phrases (never? with other patterns). The words (that is, the elements which I call words) are built with morphs, and the morphs with letters.

Predicative Patterns

Coptic sentences are formed around a verbal nucleus, obtained through the arrangement, in a certain order, of words and phrases, themselves deprieved of a "conjugation". Other "complementary" phrases are likewise obtained through the arrangement of words which may be labelled as prepositions and substantives. Adjectives and adverbs are also obtained through appropriate arrangements, and the use of some particles.

The verbal nucleus is constructed by means of one of two kinds of phrases, called respectively **bipartite and tripartite pattern**. The bipartite pattern has two possible forms, called respectively adverbial and nominal. In the adverbial bipartite pattern (**bipa**) the first part is filled by a pronominal particle or by a noun phrase; the second part by a particular class of substantives (called verbal substantives), or a "qualitative" (a class of words which comes from one form of the old Egyptian conjugation), or an adverbial phrase. In the nominal bipartite pattern (**bipn**) we distinguish two forms, one for the 1st and 2nd person (first part: personal pronoun in nominal form; second part: noun phrase), one for the 3rd person (first part: noun phrase; second part: one of the three pronouns "ⲡⲉ, ⲧⲉ, ⲛⲉ".

The **tripartite pattern** (**trip**) is formed by the conjunction of: (a) a verbal particle (last remnant of the Egyptian conjugation); (b) a pronominal particle or a noun phrase; (c) a verbal substantive (cf. above). Some of the verbal particles have a positive form, and a negative one; others are negativized by means of the particle "ⲧⲙ".

The following list formally states the situation described above. Each item is itself a system, which in particular instances may be composed of only one phenomenon.

Clauses

Some particles may be prefixed to the bipartite and tripartite patterns, to transform their meaning into past imperfect ("ⲛⲉ, ⲛⲉⲣⲉ"), circumstantial clause ("ⲉ, ⲉⲣⲉ"), relative clause ("ⲉ, ⲉⲧ, ⲉⲧⲉⲣⲉ, ⲛⲧ", etc.), and a special pattern on which the discussion is still open ("ⲉ, ⲉⲣⲉ, ⲉⲛⲧ"; so-called second tenses).

Phrases

Phrases are part of the sentence, constituted of words or other phrases, which do not consitute by themselves an utterance. According to the function that thei fill in the message contained in the utterance, we identify different categories: **nominal phrase, nominal-verbal phrase, adverbial phrase, comparative phrase**.

Terminal Categories of Words

The Coptic vocabulary is an almost unique example of blend of two preexisting ones: the

Egyptian and the Greek. In fact, the (late) Egyptian words still used in Coptic are the main component; but the Greek words, inserted in great number, cannot be properly considered "loan words". On the contrary, Coptic authors (including translators) were free to use any Greek word they deemed fit for the circumstance, according to their personal taste, as well as to traditional conventions derived from the secular contacts of the two languages. Greek nouns were used (indeclined) mainly in the nominative form, as were the adjectives, which otherwise were treated as substantives; Greek verbs assumed a simplified shape of the infinite form; the other categories (prepositions, adverbs, conjunctions) conserved their original form.

Together with the scheme of the computational grammar of Sahidic Coptic, which can be found in the web page of the *Corpus dei Manoscritti Copti Letterari* (rmcisadu.let.uniroma1.it/~cmcl) we have placed a large **list of Sahidic words**, arranged according to the terminal categories of words. In the same page we have placed some examples of the results obtained from a semi-automatic analysis of some texts, the *Historia Horsiesi*, the *Pistis Agathonikou,* the *Contra Stratonicum.* All this will illustrate better than words the shape and work of the computational grammar.

Bibliography

Robert de Beaugrande, “Reconnecting Real Language with Real Texts: Text Linguistics and Corpus Linguistics,” *International Journal of Corpus Linguistics,* 4 (1999) p. 243-259.

John McHardy Sinclair, *Corpus Linguistics at the Millennium*, Pescia, 1997.

Hartley Rogers, *Theory of Recursive Funcions and Effective Computability*, Cambridge (Mass.), 1988².