The morphosyntax of Bantu nouns

1. Outline:

In the next sections, I describe an ongoing project on nominal morphosyntax in Bantu and give some reasons why I would find it useful to establish a link with the AfrAnaph project. Our project concentrates on the morphosyntax of nominal class prefixes on nouns as well as on verbs and adjectives (the so-called concords). We also investigate the internal structure of pronouns and demonstratives, and our preliminary investigations have also led us to work on the structure of relative clauses and the structural properties of the different moods (principal vs participial vs subjunctive), since these affect the shape of subject concords.

Our goal is to uncover the abstract syntactic structures underlying the different prefixes. To this end, we study the formal correspondences between the full nominal prefixes of the different noun classes and the various concords (subject concord, object concord, adjectival concord) as well as the way subparts of a class prefix occur inside pronouns and demonstratives. To determine how class distinctions are structurally encoded, we are particulary interested in establishing patterns of syncretisms.

The project is grounded in the nanosyntactic approach to syntax as currently developed at CASTL. However, we will want our empirical findings to be accessible to researchers with different theoretical orientations.

So far, we have based our work mostly on data available in the Bantuist literature or through consultation with specialists in the field. Field work has not yet been carried out in any systematic way. But we aim at carrying out two field work expeditions per year over the next three years in collaboration with linguists at African universities. (Our contacts so far are limited to Stellenbosch, Durban and Nairobi.) If a grant application currently under assessment by the Norwegian Research Council should go through, we will also organize short summer schools for African linguistics students to train our own consultants.

Our long term objective is to expand our data base systematically by starting with a selection of the Southern Bantu languages and work our way to the North over three years.

The project is currently funded for three years by Tromsø Research Foundation (travel expenses, equipment) and the University of Tromsø (a post-doc position to be filled in the fall of 2010). It is managed by Tarald Taraldsen, and Michal Starke invests 50% of his research time to the project.

2. The project

The following provides a fairly detailed description of what we are currently doing and what we would like to do in the next couple of years.

2.1. The setting:

This project is rooted in the research carried out over the past few years at the Center for Advanced Study in Theoretical Linguistics (CASTL). On the empirical side, this research has emphasized inportance of cross-linguistic comparative study. On the theoretical side, a view of morphosyntax has crystalized that is highly decompositional and relies on a specific, non-standard conception of the relationship between morphemes and abstract morphosyntactic structure. The general framework emerging from this has come to be known as “nanosyntax”. Works such as Ramchand (2008) and a growing number of CASTL PhD dissertations bear witness to the power of the nanosyntactic world view.

Recently, an investigation of the nominal morphosyntax of a small sample of Bantu languages made us realize that this area of study may provide an ideal testing ground for the general theoretical assumptions characteristic of the nanosyntactic framework. Conversely, nanosyntax turned out to provide elegant analyses for many of the otherwise mysterious properties of the morphosyntax of Bantu nouns and agreement markers. This realization led to the project proposal we describe here.

2.2. Where we are:

We have already carried out a preliminary investigation of the properties of nominal structures in a small group of Southern Bantu languages. The results of this pilot project are reported in Taraldsen (2010). Here, we present, in condensed form, the line of analysis fo llowed so far and some of the results it has led to. In the next section, we will list some of the major research questions that arise from this investigation.

2.2.1. A technique for determining the structure of noun class markers:

The research we want to do, will exploit a new technique to discover empirically the structures underlying Bantu "class markers", and by extension the properties of the important and theoretically mysterious class of 'gender, class, case and number markers' cross-linguistically. This approach has been established by Taraldsen (2010), will be further refined in ongoing work by Taraldsen and Starke and is based on prior work on syncretism by Starke and Pavel Caha.

Consider for instance what appears at first to be a straightfoward case of class markers in Xhosa (a Bantu language of the Nguni family, spoken in South Africa, primarily the Eastern Cape):

	N	ADJ
class 2	aba	ba
class 4	imi	mi

The forms in the left column occur as noun class markers prefixed on nouns. The forms in the right column are used as agreement markers on adjectives, i.e. they are “adjectival concords” (acs) according to the traditional Bantuist terminology. What we see, is that the ac has the same shape as the second part of the corresponding nominal class prefixes, illustrating a perfectly regular pattern in Nguni.

It is mostly uncontroversial that the similarity between the markers on such nouns and adjectives in Bantu is not accidental (because it is widespread, sometimes systematic, and in some languages the two forms are fully identical), and a straightforward approach is frequently suggested: The shared part is a morpheme in itself so that 'aba' is bimorphemic 'a-ba', where the first morpheme (the “augment” or “initial vowel”) can in fact be shown to have a morphosyntactic life of its own. How to make this generic idea work in a systematic and deductive way, has however remained a mystery. The approach to morphosyntax developed at CASTL over the last few years (Starke (class lectures), Ramchand (2008), Caha (2009), Medova (2009), Abels & Muriungi (2008) , Jablonska (2007), Muriungi (2009) however offers a solution, and in fact yields a very precise calculus of Bantu class markers.

In the above example, the morphological structure of 'aba' would be:

[x A- [y -BA]]

And the structure corresponding to nominal class markers in Xhosa would be [X [Y]], while adjectival class markers would consist only of [Y]. Consider what happens now in class markers for subject agreement on verbs, the so-called “subject concords” (scs). In class 2, the sc is *ba-* and it would therefore be natural to suppose that the underlying features of class markers for subject agreement on verbs is also [Y]. But in class 4, the sc is *i-*:

	N	ac	sc
class 2	aba	ba	ba
class 4	imi	mi	i

This comes as a surprise on the idea that the scs are [Y], and this is where simple traditional approaches break down. Notice however that the class 4 sc *i-* is not entirely surprising: it consists of the [X] layer of the nominal marker. What we need to express is that subject agreement of class 2 corresponds to Y while in class 4 it corresponds to X.

The starting point of the solution is the idea that a given morpheme can be the direct realization of more than one grammatical terminal, i.e. a morpheme can realize an entire syntagm, or span of terminals. This assumption runs counter all traditional approaches to linguistics which locate morphemes/words inside terminals, and that has been the focal point of the research at the CASTL research group over the last years. It will be shown below how that idea provides a solution in a way that other approaches don't.

The second ingredient to the solution is that grammatical categories like 'noun class marker', 'subject agreement marker' and 'determiner' can be differentiated in terms of how big a chunk of the full nominal structure they correspond to. This idea is natural in a context of relatively large grammatical structures, where lexical items correspond to an entire span of the structure, and has been explored in various seminars by Michal Starke.

Given those premises, there is a straightforward solution to the puzzle of Xhosa scs: if the underling structure has three layers [X [Y [Z]]], not only two, the nominal agreement markers (the concords) may now be seen as follows,

$[_X$ a $[_Y$ ba $[_Z$ ba
$[_X$ i $[_Y$ i $[_Z$ mi

This should be read as follows: the morpheme *ba-* is a realization of both Y and Z, and the morpheme *i-* is a realisation of both X and Y. Class markers on nouns correspond to the full structure as before, and acs correspond to the lowest layer only, as before. With one proviso to be discussed below, this means that the ac will be *ba-* in class 2 and *mi-* in class 4, as needed. If the scs correspond to Y, it now follows that the sc will also be *ba-* in class 2, but *i-* in class 4, as is indeed the case.

Implicit in the reasoning above is a third, and final ingredient of the solution: associated to the idea of morphemes realizing multiple terminals, is the idea that a morpheme can realize any subset of its associated span. That is, the lexical entry for the morpheme is a superset of the grammatical structure it lexicalizes. For instance, the morpheme 'ba' realizes both Y and Z in nominal class markers, thus corresponding to its complete span, but in adjectives it only realizes Z and in subject agreement it only realizes Y, both of which are proper subsets of its full lexical span YZ.

In his preliminary report on the matter, Taraldsen (2010) shows that entire system of the Nguni class markers, including its more esoteric aspects such as the demonstrative system, follows entirely from such a system. For example, the analysis captures the properties of the distinct set of object agreement markers on verbs (the “object concords” – ocs) by adding another structural layer intermediate between Y and Z, giving the following representation for the full structure underlying the class prefixes on nouns, with more mnemonic labels for the different layers:

$[_{\text{Augment-layer}}\ X\ [_{\text{SC-layer}}\ Y\ [_{\text{OC-layer}}\ W\ [_{\text{AC-layer}}\ Z$

The system also provides a straightforward characterization of the fact that Swati nouns, unlike Xhosa or Zulu nouns, have an initial vowel only in the so-called nasal classes, e.g. in class 4, but not in class2. This follows if Swati nouns do not project the Augment-layer. This system is precise enough that given a full table aligning full nominal class markers, the different class agreement markers and demonstratives, one can in fact algorithmically derive the precise structural decomposition of each category. It is also restrictive - it cannot accomodate just any data. The prediction made is that the paradigms of class markers will be linearizable into spans, such that cutting the spans will yield the surface shapes. This is a generalization known as the *ABA elsewhere in the morphosyntactic literature (Bobaljik (2007)) and may well be falsified. So far, the languages we have had access to fall into the prediction.

2.2.2. Prospects:

So for the first time, we have a precise calculus of class/agreement markers. This opens up new possibilities of research into the traditionally mysterious area of 'class/gender' marking. In particular, this methodology can be applied to any Bantu language with a sufficiently rich class paradigm in order to obtain comparative data for the underlying structure of class markers in many related languages. Doing that will provide us with rich and very precise data delimiting the range of possible underlying structures of class markers. This data will then be precious both for typological purposes and for theoretical purposes.

In this project, we intend to pursue this new opportunity and make discoveries about class markers and, more generally, the structure of noun phrases through a systematic extension of this reasoning in a comparative study of Bantu nominal morphosyntax.

2.3. Moving onwards:

Our investigations will be guided by a number of research questions emanating from the pilot project (Taraldsen (2010)). We now look at a selection of these. In each case, we indicate what kind of empirical work is needed to settle the issue and this will in part determine the research agenda presented in section 4.

2.3.1. The content of the nominal functional heads:

The analysis sketched in section 2 ultimately leads to the conclusion that the Nguni nominal prefixes correspond to at least 4 distinct functional heads organized into a hierachical structure. This, however, invites the question how these heads are to be characterized in terms of their syntactic and semantic content.

While the higher layers of nominal structure are plausibly D(eterminer)-related (judging in particular from the distribution of "initial vowel drop"), the lower heads relate to the determination of number and gender.

Number must be determined by a constellation of different heads giving rise to different types of sg and pl readings. In Northern Sotho, for example, a class 9 noun like *kgômo* "a head of cattle" can both form the expected pl *dikgômo* in class 10 and *makgômo* in class 6. But there is a semantic difference between the two pl forms. While *dikgômo* means "cattle", *makgômo* means "herds of cattle" suggesting pluralization on top of an operator forming groups from individuals.

Gender, or "noun class", is also, we will argue, determined by sets of semantically loaded heads rather than by class diacritics. The latter approach would cut us off from accounting for partial syncretisms such as the one between the class 8 prefix *–zi-* (Swati *–ti-*) and the class 10 prefix *–zi-N-* (Swati *–ti-N-*) in Nguni. Nor would it provide a basis for understanding the partially fixed pairing of sg and pl classes as well as the non-random exceptions to the pairing, such as the one mentioned above for Northern Sotho.

Using regular syntactic/semantic features, we are able to capture the partial syncretisms by characterizing two classes as sharing the same features in a specific structural region while having different features in some other part of their structure. The fixed sg/pl pairing, on the other hand, can be expressed by assigning the paired classes identical features except for the number-related ones. To capture the systematic exceptions to the fixed sg/pl pairing, we will have to study the interaction between things like the group-forming operator of Northern Sotho and the gender-related features.

While assuming that genders/noun classes are to be characterized in terms of configurations of syntactic/semantic features, seems well motivated, it is less clear how to determine the exact semantic content of each feature, even though the choice of class prefix seems to affect the meaning in a well-known range of cases. Therefore, we will side-step this issue in the initial phase of our work, limiting ourselves to tracking the distribution of features across noun classes by looking at syncretisms and sg/pl pairings. For this to be sufficiently informative, we will need to study the relevant facts in a reasonably large sample of Bantu languages moving beyond the limits of Nguni.

2.3.2. The structure of the agreement markers:

As seen in section 2, the set of Nguni scs is not identical to the set of acs, and we must identify the two categories with different layers of the structures underlying the corresponding noun class prefixes on nouns. This analysis extends to the ocs, a set that coincides neither with the scs nor with the acs.

A sc contains structure from tthe SC-layer. An oc starts at the top of a lower layer, the OC-layer, and an ac contains only the heads in the lowest structural layer, the AC-layer. Giving a theoretical interpretation of this is an interesting challenge.

A line of analysis worth pursuing would start from an exploration of the relation between the different heads forming the structure of nominal prefixes and the lexical classes noun, verb and adjective in Bantu. Grounding this in work emanating from Cinque (1999), we may take it that all syntactic ("functional") heads are hierarchically ordered in a strict "functional sequence" (fseq). If we also assume that the same fseq unfolds on top of the root in any lexical class, we may account for the size difference between scs and ocs by taking them to attach in different positions within the fseq projected on top of the verb, i.e. the oc would attach in the OC-layer of the verb while the sc would attach in the SC-layer of the verbal projection of the fseq. This would correlate both with the higher position of the sc and with the different properties that seem to converge on identifying the part of the verb which the oc attaches to, as a proper substructure of the full verbal form (the so-called “macrostem”). The acs would attach even lower, in an area of the fseq (the AC-layer) which is lexicalized by the root itself in verbs, but not in nouns or adjectives, in conformity with Baker’s (2005) proposal that a V contains an A in its structure.

There is also an important empirical question. We want to find out whether the pattern ac < oc < sc is replicated across Bantu or only in a typologically definable subset of the Bantu languages. The outcome of this investigation will inform our implementation of the analysis just sketched (or lead us to abandon it).

2.3.3. *ABA:

As already explained, our analysis relies on a non-standard theory of the relationship between syntactic heads and morphemes: The lexicon associates a morpheme with a possibly non-trivial span of heads, and a morpheme can lexicalize any subset of the span associated with it. Under this theory, lexicalization is governed by two supplementary principles conceptually related to the familiar “elsewhere principle”.

Given the structures assigned to the different concords in Taraldsen (2010), we can show that this theory of lexicalization correctly predicts that while the languages in our sample show the syncretisms sc = oc = ac, sc = oc ≠ ac and sc ≠ oc = ac, we don’t see sc = ac ≠ oc, an instance of the illicit ABA-pattern under our analysis.

However, we do not yet know whether the sc –oc –ac series never gives rise to ABA syncretisms in Bantu outside Nguni. We have so far looked at Northern Sotho and Venda, which replicate Nguni, and have started investigating Kitharaka, but need a much more extensive survey to able to say that the concords never syncretize as sc = ac ≠ oc. This survey will go hand in hand with the empirical investigation mentioned at the end of 3.2.

2.3.4. Bantu languages where nouns don’t have initial vowels:

Adopting the theory of lexicalization just sketched also leads us to a controversial conclusion when we extend our account to Bantu languages outside the Nguni group. In Northern Sotho, for example, nouns never have initial vowels, e.g. class 1 *mo-tho* “man” corresponds to Xhosa ***u**-m-ntu*. A priori, we could have chosen to say that the syntactic structures are exactly the same by allowing the morphemes lexicalizing in lower layers to extend their lexicalization domain upwards in Northern Sotho, i.e. the heads lexicalized by *u*- in Xhosa *u-m-ntu* would be lexicalizd instead by *mo*- on Northern Sotho. Crucially, however, the Northern Sotho scs are just like their Nguni counterparts, e.g. sc1 = *u*-. Since our theory of lexicalization will not allow heads in the SC-layer to lexicalize as *mo*- inside the nominal prefix, but as *u*- in a sc, without losing the account of the inexistent ABA-pattern mentiond in 3.3, we are ultimately forced to conclude that nominal structure doesn’t project beyond the AC-layer on top of nouns in Northern Sotho, i.e. Northern Sotho noun phrases lack the D-related heads. This feeds into the general theoretical discussion as to whether argument noun phrases must be DPs or not, and also leads to an examination of the parameters that differentiate languages like Northern Sotho from Nguni.

On the empirical side, we therefore want to carry out a fairly extensive survey to determine which other properties (if any) correlate systematically with the presence/absence of initial vowels in Bantu languages. These properties could be found in the DP-internal syntax (positioning of demonstratives and modifiers, for example) or at the clause level.

We also need to accommodate Bantu languages that show initial vowels in some classes, but not in others. In fact, there is one such language in the Nguni group. As already mentioned, Swati has an initial vowel only in the nasal classes, a situation which is characterized in Taraldsen (2009) by saying that Swati nouns do not project functional structure beyond the SC-layer, leaving out the Aug. This suggests that there might also be Bantu languages intermediate between Swati and Northern Sotho, i.e. languages in which nouns project to the top of the OC-layer, but not further. Such a language would be like Swati except that it would not have an initial vowel in class 1, everything else being equal.

If the Bantu languages really form a continuum with respect to the distribution of initial vowels over the noun classes, and the continuum is generated by shrinking nominal projections in the way described here, we expect variation to have a certain profile. For example, there should be no Bantu language with an initial vowel in the non-nasal classes, but without initial vowels in the nasal classes. In the course of our empirical investigation, we want to find out whether this expectation is actually fulfilled.

Also, if we succeed in establishing systematic correlations between the absence of a higher structural layer on top of nouns and the existence of other properties of either a language's DP-internal or clause-level morphosyntax, we should expect to see sets of such properties forming a chain of proper inclusion relations following the continuum of languages determined by the distribution of initial vowels. For example, the set of properties linked to the absence of the Augment-layer in Swati should be properly included in the set of properties linked to the absence of both the Augment-layer, the SC-layer and the OC-layer in Northern Sotho. Determining whether this actually holds, is another important goal of our empirical investigation.

2.3.5. Summary:

Based on what we have learnt from the study of Nguni, we have developed an analytical strategy and started approaching an understanding of the Bantu data. We feel justified in thinking that pursuing the research goals set out above will deepen this understanding and will both contribute to a more accurate description of Bantu nominals and the Bantu concord system and put a variety of interesting theoretical hypotheses to the test. We would also like to emphasize that although nanosyntax is a new approach to morphosyntax, it has already, as mentioned in 2.2.1., led to significant results, and it therefore seems eminently reasonable that one should want to develop it further by applying its basic principles in a larger empirical domain.

2.4. Implementation:

We now turn to how we want to implement this project. First, we describe the research strategy we intend to follow. Then, we specify our manpower needs Finally, we set up a time table, indicating the major mile stones.

2.4.1. The research strategy:

We can obviously not hope to be able to investigate every Bantu language in depth. Nevertheless we think we can build our analyses on solid empirical ground by adopting the "selective global comparison" strategy pursued at CASTL. Consequently, we will study three different subgroups of Bantu in great detail, and supplement this research with information gleaned from the descriptive literature on other Bantu varieties.

The first component of this strategy entails field work in three different areas of the Bantu speaking region and close collaboration with local linguists. Initially, we intend to set up a research network with nodes in Stellenbosch (for Nguni), Durban (for Zulu and the non-Nguni Bantu languages spoken in South Africa) and Nairobi, where our former student Peter Muriungi now is a university teacher. Our collaborators in these locations will assist in individuating particularly relevant language varieties in their areas and will be instrumental in identifying suitable native language consultants and field researchers, who we aim at recruiting from their graduate students. These should establish local networks around the three main nodes and be available as "fact finders" throughout the project period.

The Tromsø researchers will also use these networks in their own field work in the Bantu-speaking area. As we learn more, we will expand the network to include other locations. Taraldsen and Starke have already worked in South Africa with linguists from Stellenbosch and Durban, where they also gave talks discussing nanosyntactic issues in February 2010. Using funding from Tromsø Forskningsstiftelse, they intend to continue this activity in the academic year 2010/2012.

To be able to use graduate students from Stellenbosch and other African universities as field researchers, we need to introduce them to field work techniques wherever the local universities do not provide field methods courses. We plan to achieve this by organizing short summer schools (two weeks) in different locations in South Africa and Kenya, drawing in part on the field methods expertise developed in the NORMS and ScandDiaSyn projects at the University of Tromsø and purchasing technology developed at the Text Laboratory of the University of Oslo for data storage.

Since we also want to enhance the participating students' awareness of the theoretical issues involved, these summer schools will also offer classes on a selection of topics in current morphosyntactic theory taught by us and colleagues from Tromsø and other universities. Thus, the summer schools will complement the linguistics programs available at African universities and be a vehicle for bringing new theoretical developments to the attention of African graduate students. Ultimately, our goal is to enable at least some of the graduate students working as field researchers to become active participants also on the theoretical side.

The second component of the selective global comparison methodology requires access to written documentation. There is a large body of literature on Bantu languages, but much of it is not electronically available or available in the Tromsø University library. Primarily, we want to do research in the libraries of Leiden University, Stellenbosch University, the University of KwaZulu-Natal and the University of Pretoria.

3. Collaboration with AfrAnaph

The research strategy described above involves a significant amount of traditional field work. Since field work is time consuming and labor intensive, we will only be able to cover a selected few of the Bantu languages with this strategy. Yet, it is crucial to obtain reliable empirical coverage across Bantu. In part, we can ensure this by relying on available descriptive work. But even the best descriptions are bound to leave out details that may well be crucial to a researcher approaching the data from a specific theoretical angle. So, using electronic questionnaires seems likely to become an essential ingredient in our empirical work. Since it is difficult to use this approach with untrained native consultants, it is natural for us to seek access to a pool of consultants who are themselves linguists. This is where our needs may intersect with AfrAnaph's needs.

What we can offer, is access to a field of study currently outside the purview of the AfrAnaph project and yet related to the study of anaphoric expressions in numerous ways (for example, the structure of pronouns and anaphoric vs non-apanhoric object concords). Our own research also leads us into areas independent from nominal morphosyntax but yet relevant to it as well as to almost any other syntactic topic in African linguistics, e.g. the relevance of syntactic structure to the determination of domains for morphophonological processes (tone spreading, vowel-contraction vs glide-insertion etc) Sufficiently detailed information bearing on these issues should be of value to any project under the AfrAnaph umbrella, and can most easily be obtained in collaboration with native consultants trained in linguistics.