NP Analysis in OpenLogos

B. E. Scott April 14th, 2010

Some General Observations:

1. OpenLogos was designed primarily to translate commercial and technical documentation. NP's in technical documentation particularly can be quite complex, e.g.:
   "The 2-Part Version 5 Checker Program CP5"
2. Rendering a NP like this in another language can be daunting task for any MT system (I doubt OpenLogos can handle it, actually). But attempts to handle complexities like this have caused target handling of NP's in OpenLogos to itself become extremely complex.
3. Adding to the complexity is the fact that target work is done in so-called 30, 40, and 50 tables, diagnostics of which are not always easy for the beginner to understand. For your information:
   - Optional 30-tables are tried to individual source rules (assuming the source rule has target implications).
   - 40-tables are sharable (i.e.common) target functions callable by 30-tables
   - 50-tables are <u>interlingual</u> sharable target functions (i.e., callable and suitable for any target language). These are callable by 30- or 40-tables.
   - Note: If you have access to the tables in their original data form (before rule generation), one should hopefully find comments by the rule-writer that may help understanding the functions performed.
4. From 1970-1988, the Logos system did not employ target tables.
   - All target functions (switches) were included as part of the action (VTR) part of a rule (VTR = vector transform). While most switches are made either for source or target purposes, these actions (switches) originally co-existed side by side.
     a. In 1988, separate target tables containing target actions were introduced to make the system multi-target. In effect, a single body of <u>source</u> rules could now be employed for any number of targets.
        - I.e., a source rule could call target tables for any number of target languages.

     b. Interestingly, however, if one were only interested in a single target (e.g., English-Hindi), then these target tables could conceivably be dispensed with, although one would have to be aware of VTR length limitations.
        - Note that the second rule in the discussion below reflects this old practice of mixing source and target actions in the VTR.

     c. In either case, placing sharable code in 40- and 50-tables obviously is more efficient.

Overview of TRAN1 processing of the NP:
The tall handsome boy

This overview will serve as Introduction to the commentary placed directly inside the TRAN1 diagnostics for this simple NP (diagnostics are provided separately).   (TRANs are sometimes referred to as PARSEs in some system documentation, in which case target actions are referred to as TRANs).

- In what follows, we identify the source rules that sequentially match on this simple noun phrase in the course of parsing it as NP, and we provide a brief overview of both source and target actions.
- We will not here go into the details of the source and target operations, e.g., setting of particular cells and other communication areas. Such details can be viewed in the diagnostics provided separately.

```
***** A MATCH STARTING AT  1 LEVEL  1
**809 BOS = -1 / T90,F02/CK FOR ? (ADD VC108 FOR ADV-S287) ST286 EGSP1
 SP LINE:  1 (20 1 1)
 VTR: -42  10 809   1   1 -20   0 108   0 -55  19 -81  62 -55  70   1   0
      -55  99 1   0 -46 -81   0   900   2 -41   1 999
```

- This is a BOS rule (BOS=Beginning of Sentence). Wc20 is used for all punctuation. Type 1 is unique to BOS.
- Source Actions:
    - i. Rule branches to a set of nested rules in wc10 of the rule base (i.e., pattern dictionary) (The TRAN rule bases (one for each level of TRAN) are organized (sorted) by SAL word class.
        1. The type code 809 is merely a wc10 rule address, not an actual SAL code.
    - ii. Rule backspaces all the way, hence no SWORK is formed by this rule
        1. SWORKS are formed automatically from the last element of the SP line unless the system is otherwise instructed by some switch (e.g, -20 sw).
    - iii. Various cells are set and the SAL type field of BOS is set to 900.
    - iv. No use is made of this 900 in subsequent matches here but potentially it would enable a subsequent match had the pattern input pattern called for it.
        1. Altering any of the wc/type/form field of an element is a principle means rule-writers have for influencing subsequent rule matches.
- Target Action.
    - i. None

```
***** A MATCH STARTING AT   1 LEVEL
PUNC = PUNC ST1184 EGSP1
SP LINE:  1  (20 -1 -1)  (BOS)
VTR: 83   0  -1   0  84   0 999
```

This rule processes any punctuation.

- Source Action:
    - SWORK is automatically formed in the absence of any switch inhibiting it.
        - Consistent with this, there is no backspace in this rule (see below for further discussion about backspacing).
    - Unless otherwise specified, default action is to form SWORK from last element of the SP line.
- Target Action:
    - OPADR (target output address array)is created for this punctuation constituent. (BOS is looked upon as a punctuation)
        - Each source constituent (e.g., NP) has its own OPADR.
    - Contents of slots (83 and 84) are unloaded and target address of relational pointer (-1) is placed into OPADR. These things are placed in order of occurrence.
        - Relational pointers point to elements in the SP line
            - Relational pointers in VTR’s range from -1 to -10.
                - A rule can have from 1 to 10 elements in the SP line.

- Relational pointers within switches range from -81 to -90
- Note: Slots are receptors or containers into which target elements and functions can be placed by earlier rules, pending unloading into OPADRs.
  - Slots may contain switches (functions) as well as target addresses of elements.
    - Switches are executed at unload time
    - Elements (address of target transfers) are loaded directly into OPADR
- See notes below for further discussion about OPADRs and their formation.

```
***** A MATCH STARTING AT  2 LEVEL
DET .S(ADV). ADJ .S(ADV). ADJ .S. N EL(NON-N)=-A*0/F13=ADJ S1286ESM987 CMG9/91
SP LINE: 8  (14 -1 3)  (55 -1 -1)  (1 -2 47)
    (55 -2 -1)  (1 -2 -1)  (58 -2 -1)  (1 -1 80)  (-1 -2 -1)
VTR: -20   0
   -22   2 -83   1  -3   0  -7  69 -46 -83   0 851   0 -36  56   0
   -63   1 770   1 -41 100 999
```

This rule is uncharacteristically lengthy.
-3 in form field of article (wc14) signifies singular OR plural morphology
The -2 in the type fields in the SP line denote tagsets which appear immediately below the SP line, in order of occurrence.

- Tagsets have three purposes: (1) constraint specifications for matching a given element; (2) expansion of SAL types on which the rule may match; (3) for stretch (.S.) elements, specification of the kinds of SWORKs a stretch element may stretch over.
- Stretch elements are like Kleene stars.
  - In the above SP line, wc 55 and wc 58 are stretch elements.
- Source Action:
  - Rule backspaces all the way (i.e., does not produce an SWORK)
    - A*0 expresses this total backspace. It reads: backspace all but 0. This is the convention for backspacing where stretch elements are involved (i.e., where the rule-writer does not know the length of the input pattern actually being matched)

    -41 100 is switch and parameter setting for backspacing all the way in context of a stretch.
  - -22 switch sends the first adjective ‘tall’ (-3) to the Semantic Table (SEMTAB) in the context of the presumed head noun (-7, i.e., 1 -1 80)
    - Note: Actions in SEMTAB can be for source analysis or target transfer, or both.
    - SEMTAB does not have 30-tables.
  - The SAL type of the adjective ‘tall’ is re-labled 851, telling subsequent rules that this adjective has been to SEMTAB.
  - Calls 30-table #1770 for target action
- Target Action:
  - 30-table calls 40-table for common target action.
    - 40-table tests various communication areas (cells and scons)
    - A value is set in scon 20 (reserved for targets) signifying “no combining form” is needed in conjunction with German adjective transfer.
    - no match occurs on the ‘Adj + head noun’ send to SEMTAB, i.e. the original dictionary transfer will be used.
    - no other significant target action.

```
***** A MATCH STARTING AT   2 LEVEL   1
DET = DET E1 ST282 MMT287
```

(E1 here probably means English TRAN1. The rest are linguists initials and dates)

SP LINE: 1 (14 -1 -1)
VTR: -20   0 -31 15
-63 0 458 1 999

- Source Action:
    - SWORK formation is suppressed (-20 sw)
    - ARTDEF (definite article) flag is turned on (-31 15). This will influence form field of the SWORK that will shortly be formed for the NP.
    - 30-Table #458 is called
- Target Action:
    - A pointer to the target transfer for ‘the’ in loaded into a VC.
        - Note: VCs are empty, therefore variable, constants into which pointers to target transfers can be inserted.
        - This VC will later be loaded (by the target action of a later rule)into a target output address array called OPADR.
        - VCs can be accessed by rule-writers even after they are loaded into OPADR, which is a key reason for having them.

GENERAL NOTE

From this point on, actions in target tables will be loading target transfer pointers into slots or VC’s for eventual unloading into the OPADR.

- The eventual building of the target OPADR for the current NP will occur when head noun ‘boy’ is processed
- It will help to keep in mind the German target template by which all German NPs are formed (shown below). (Note: Each target language has its own template conventions.) We present some general information first, applicable to all target languages.
    - Numbers 71 to 98 represent slots. Slots are receptors that are filled by earlier rules and that get emptied at time head noun is processed.
        - Slots may contain:
            - pointers to transfers of elements in the SP line
            - constants (pointers to German words e.g., which SEMTAB rules may have substituted for the original dictionary transfers
            - VC’s
            - Any number of Switches which execute other target actions. These switches will be executed when slot is unloaded.
    - Numbers from 101 120 to represent VCs. VCs are receptors of a more limited kind, that may be addressed once they are loaded into the OPADR.
    - Numbers from 131 to 9998 are target dictionary constants. Constants are addresses of target words in a target-specific dictionary (separate from the main dictionary) that the rule-writer wishes to use in target generation.
        - When SEMTAB alters an original dictionary target transfer to some other word, such new words come from this so-called constant dictionary.\
            - Only SEMTAB rules allow for 4-digit constants and other values.
            - Note that the higher frequency constants lie in the lower (3-gidit) register of the constant dictionary and hence can be used in 30- and 40-tables (which limit all values to 3 digits).
        - N.B. numbers from 121 to 130 are interlingual signs such as punctuation, etc.
    - Note: Some of the ranges expressed above may be not 100% correct (i.e., off by a number or two.), as I am relying on memory.

GERMAN NP TEMPLATE

102 75 107 83 73 103 105 108 106 71 77 101 HEAD NOUN 104 74 112 79 96 110

The linguists responsible for a given target language set this template up amongst themselves at the beginning. They seek to provide target language equivalents for every conceivable (simple) NP in the source language.

- Each VC or Slot represents a particular kind of element that conceivably can be encountered during analysis of source noun phrases, and that therefore must be provided for in the target equivalent.
- Preparation for the target NP is accomplished <u>incrementally</u> by loading slots during source analysis of the NP with VC’s and target transfers.
- When the Head Noun of the source phrase is finally processed, the slots are unloaded into the OPADR in the order indicated above, thus insuring proper target language generation (which takes place at the end of TRAN4).

```
***** A MATCH STARTING AT   3 LEVEL   5  (tall handsome boy,)
ADJ1 ADJ2(EXCL F6) .S.N EL(NOT N) = (ADJ ADJ) -A*1 BES0287 ESM0387
SP LINE   5  (1 851 23)  (1 -2 47)  (58 -2 -1)
             (1 -1 -1)  (-1 -2 -1)
VTR -22   2 -82   1  -2   0  -4  69
     -63   3 159   1 -41 101 999
```

Comment line incorrectly suggests that both adjectives are disposed of. (ADJ ADJ) should read (ADJ)

- Form 23 denotes pure descriptive Adj.
    - Form 23 differentiates adjectives from nouns, both of which share wc 1.
- 851 signifies Adj has already been to SEMTAB
- <u>Source Action:</u>
    - Second adjective (‘handsome’, i.e., 1 -2 47) + head noun is sent to SEMTAB
    - Backspace is all but 1 (which tells us that (Adj Adj) was incorrect (may have been true at once time however)
    - Calls 30-table # 3159
- <u>Target Action:</u>
    - 30-table calls 40- and 50-tables which set various cells and tests various cells and scons
    - 50-table 71 loads into Slot 73 the following: slot 83, slot 71 and a pointer to the transfer for ‘tall’
        - Note that slots can contain slots.
        - As stated, these slots will be unloaded at Head Noun processing time.
        - It appears that the 50-table here may possibly <u>not</u> be interlingual. Not sure of this however. (No target word order is as yet established)

```
***** A MATCH STARTING AT   4 LEVEL   2
**108 N N = -2 / SMTB(69 IN SEND -02069) E1 ESM0387/89 MMT1288
 SP LINE  2  (1 -1 45)  (1 -1 80)  (handsome boy)
 VTR  -42  10 108   1   2 -20   0
    -22 902 -81   1  -1  75  -2  69 -54   1 -82  42   1
    -66 299 199 -82    62 852 777 -81 262 852  60
    -57   1 -46 -81   0 851   0
    -63   4  16   3
```

```
   -57   2 -36  56   0 -41    2 999
```

- Source Action:
    - Branch to wc 10 rule # 108 for nested match (see rule immediately below). Valid match occurs and its VTR (below) is processed in lieu of the VTR above.

```
02 ***108 ADJ(23) N = -2 / SMTB / 4 (VERY) PROTECTIVE N E1 CMG3/89 BMO287
  3  (10 108 1)  (1 -1 23)  (1 -1 -1)

-20   0
-22   2 -81   1  -1  75  -2  69
-63   4  21   3 -46 -81   0 851   0 -31  56 -41   2 999
```

    - SWORK formation is suppressed (-20 sw)
    - ‘Adj + head noun’ is sent to SEMTAB. No match occurs
    - Adjective ‘handsome’ (-81) is given type code 851 to indicate it was sent to SEMTAB (-46 switch)
    - Target 40-table 4021 is called.
    - -31 56 tells system not to re-match on this rule (given total backspace, i.e. no looping))
    - Backspace is all the way.
- Target Action:
    - Target table performs various tests and sets a target language communication area (scon 20) to tell a subsequent rule that no agglutination of 1st element with the head noun is required.
        - (recall that this is German target where NP agglutination is common)

```
***** A MATCH STARTING AT   4 LEVEL   1
ADJ = (ADJ) E1 ST586 MMT0288
 1  (1 851 23)  (handsome)
   -20   0 -55  34 -81  11 -55  35 -81   2 -31  36
   -63   0 161   1 999
```

In comment line, = (Adj) tells us that the adjective is being sent to some receptor in the NP template. The parens indicates that no SWORK is being formed

Type 851 specifies Adj has already been to SEMTAB

- Source Action:
    - -20 sw inhibits SWORK formation.
    - Cells are assigned values for benefit of subsequent rules
    - Target 30-table 161 is called.
- Target Action: (basically same as for the adjective ‘tall’, above)
    - 30-table 161 calls 40- and 50- tables which set various cells and tests various cells and scons
    - 50-table 71 loads into Slot 73 the following: slot 83, slot 71 and a pointer to the transfer for ‘handsome.’
        - As noted earlier, slots can contain slots.
        - These slots will be unloaded at Head Noun time.
        - It appears that the 50-table here may possibly not be interlingual. Not sure of this however. (No target word order is as yet established)

```
***** A MATCH STARTING AT   5 LEVEL   2
N(91) NON-N = -2 / CK S3 STS586 EGSP1
2  (1 -1 91)  (-3 -1 -1)  (boy + non-n element)
   -20   0
   -63   1 352   3 -36  56   0 -41   2 999
```

91 is a superform which includes all form codes for wc 1 that pertain to nouns (i.e. excludes adjectival forms)

-3 is a super-wc which includes all wc’s except wc1. It is used here as a NP delimiter.

- Source Action:

    - SWORK formation is inhibited (-20 sw)
    - Backspace is all the way
    - 30-table 1352 is called.
    - -36 sw 56 prohibits re-match on this rule (necessary, given the backspace all the way and the fact that no changes ere made to wc/type/form)
        - Note: -36 56 duplicates anti-loop function of -31 56 (in previous rule)
- Target Action:
    - 30-table sets cells and scons and tests for certain conditions.

```
**107 N = N / CK FOR F20,39,PN
1  (1 -1 -1) (boy)
   -42  10 107   1   1
   -63   0 802   1 999
```

This is the key rule that processes the head noun (‘boy’) and that causes all address pointers to be loaded into the OPADR, in the order specified by the rule-writer in this rule. Note that the lack of parens in the comment line indicates the head noun is loaded into the OPADR

This rule also causes an SWORK to be formed for the NP constituent. The SWORK will have the wc/type/form of the head noun, except as possibly modified by this or some preceding rule. E.g., the form field of the SWORK created for this NP (viz., 17) will indicate NP is singular and has a definite article.

- Source Action:
    - Branches to wc 10 rule # 107 for possible match. Unsuccessful
    - Checks various communication areas (cells) for various conditionse. E.g., is noun a process noun (PN).
    - 30-table 802 is called
- Target Action:
    - 30-table 802 tests various cells (e.g., for ING form) then initiates a complicated series of target table calls, outlined here:
        - 30T 802 calls, successively, 40T 138, 40T 139, 40T 41
        - 40T 138 sets cells
        - 40T 139 is the work horse. It calls, successively, 40T 7, 40T 39, 40T 45 for checking and testing, then:
        - Control is returned from T40 45 back to 40T 139 which, after testing various communication areas (cells, scons), proceeds to empty slots and load the OPADR with target address pointers and with VC’s (See this OPADR below).

        - The principle work of NP formation, both for source and target, is performed in 40-table 139 by the -25 switch.
            - -25 switch identifies noun as a head noun. This causes the following:
                - SWORK is formed from we/type/form of head noun, unless altered by some switch.
                - All modifying elements of NP (e.g., in this case, article and adjectives) are made to agree with head noun with respect to number/gender/case
                    - (Note: case can be specified by the -25 switch’s single parameter. among other ways)
                - Note: -25 sw is one of a few switches that have both source and target functions. It is one of the earlier switches in Logos development history
        - Control now returns to 40T 138 which finally calls 40T 41

- 40T 41 re-initializes a host of cells to zero.

Below is displayed the SWORK and associated OPADR formed by the above rule:

```
SWORKO =     1    70    17     5   (boy)                  3    16
OPADRO =    -102  -107  -107     3     4  -103  -105  -108  -106  -101
             5  -104  -112  -110
```

Note: The positive numbers in the OPADR are pointers to where the target transfers for English words are found (e.g. **5** -> Junge).

This string of addresses and VC’s constitutes the OPADR entity associated with the SWORK for ‘boy’. Such OPADR entities can be moved in relation to other constituents of the sentence, in accord with necessary target language style. The case setting for all unlocked elements inside this OPADR can be changed by subsequent rules (in higher levels of TRAN, typically).

Transfer for the head element in this OPADR entity (e.g., boy) can be modified by subsequent rules (provided it has not been locked). non-If a head element change by SEMTAB affects number and gender, then these values for the determiners, adjectives, will automatically be modified. However, transfers of elements interior to a NP cannot be changed unless altered unless these transfers have been placed inside a VC. In the above, VC 107 contains the German article ‘Der’ whose transfer can thus can be altered by subsequent rules.

In this particular NP and associated OPADR, there does not happen to be any constants (addresses of words in a special target language dictionary). Constants are seen as positive numbers from 130-9998 (I am not completely sure if the lower number is exactly correct.)

- Note: For agglutinated target languages, rules can insert a special constant into the OPADR that points to an agglutination symbol. The target generation program that processes the OPADR will then agglutinate the words on either side of the symbol.

Some Final Remarks

1. The principle that guided development of Logos MT is simple to state: Natural language is too complex and irregular for it to be treated as a formal object, i.e., for it to be handled by a strict, logical procedure. Accordingly, a declarative system was built that relied upon rules comprising abstract patterns (SAL patterns) which could be matched against a SAL input pattern.
   - Both input stream and rulebases are expressed in SAL, such that rule matching is very much like dictionary lookup.

    - Just as in dictionary look up, matching is dictated by the input stream, not some overarching logic. Hence we often say that, in Logos, the sentence becomes the algorithm.

2. True, the use of numbers in Logos rather than words or alpha symbols makes it difficult to follow the process. Historically, numbers were employed because of Fortran (the original programming language for the system.)
    - But numbers offered pronounced advantages of efficiency as well. For example: A single number could stand for a great deal of information. E.g. the numerical pattern 1 720 2 stood for a plural noun with the semantics of "functional device."
    - Thus, words with code 1 720 2 (e.g., razors, drills, etc.) could be matched upon by a rule whose SP line had any of the following codes:
        - 1 720 02 (plural morphology)
        - 1 720 -1 (-1 = any morphology)
        - 1 34 -1 (the SAL set code for 720)
        - 1 3 -1 (the SAL superset code for 34)
    - In matching, the input stream (SWORK pattern) is used as a search argument against the rule base (a pattern dictionary, essentially).
        - Matching takes place on SAL subsets first, then sets, then supersets, and finally on universal sets (expressed by a -1 in the type field).
        - As a result of this organization, we enjoy the following benefits:
            - Rule matching always favors the semantically more specific rule first.
            - Rules that have no SAL correspondence to an input pattern are never looked at.

3. Another key advantage of numbers:
    - Rules are self-organizing (no question where they belong in the pattern dictionary (rule base. I.e., they sort themselves)

    - Linguists know exactly where to look for a particular rule.

4. Another point:
    - Rules are extremely shallow, i.e. limited in their scope and action. You will notice that the system does not allow rules to have general purpose programming code. So once one understands the numbers and strategies, virtually nothing that is strange or arcane will be found in these rules.
    - Where some MT systems might have powerful rules that are very long and complex, Logos achieves powerful effects through rules working in combination with each other, executed in sequence.
        - The rationale for this is simply that this arrangement has made it extremely easy for linguists (computational or otherwise) to maintain the rule bases. They know exactly where to look for a rule, where and under what conditions it may fire, etc.
    - A final advantage of this arrangement is that the rule bases can grow arbitrarily large, like dictionaries, without significant impact on system performance.
        - This explains why Logos developers claimed that the system in principle has no theoretical limit to its improvement.

It may be of interest to know that an effort was made in the last years of Logos Corporation’s life to get rid of numbers, i.e., convert numbers to alpha symbols. The effort did not bear fruit,but there does now exist alpha symbols for SAL codes, available at http://logossystemarchives.homestead.com

****