A full example

Suppose we want to identify all the references to particular military units within some document, e.g. “ninth Army”, “24th MEU”. To do this we need to create a set of parsing rules for Sparser to apply, a representation for military units for these rules to use to store what they have found, a way of indexing the elements of this representation so that every reference to, e.g., the ‘2nd support command’, is taken to the same instance of a military unit rather than creating a new one, and a way of signaling that military units are ‘interesting’ and should be included in objects that are extracted from the document.

Analysis

The first step in developing a representation and rule set for Sparser is to look closely at the way references to the things of interest are phrased. For things like military units, the conventions of the US military in how units are to be named are comparatively easy to analyze. For the simple cases, the name of a unit is composed of its ‘type’ – “Army”, “Division”, etc. – combined with a number.

The next consideration is whether these two parts have already been modeled, in which case we can just reuse them, e.g. ordinal numbers like 24th, or whether it’s likely that they will useful in other relationships, in which case we should define them as separate notions now in anticipation of that later use. (Compare one of the ways that we describe something’s age: “42 years old”. The number and the unit of time are eminently reusable, but this use of “old” is idiosyncratic and stays within the rule system.)

Categories

The different types or kinds of objects that are modeled in Sparser are represented by ‘categories’ (see §9.1). Based on the simple analysis, we will want one category for the different types of military units, and another category for military units per se, which will incorporate a reference to the unit type along with its distinguishing information. For the moment, let us just look at how we could define the category for ‘type-of-military-unit’ and how it leads to the automatic creation of parsing rules, and take up the representation of full military units later on.

Categories are created by evaluating a special form as described in detail in §9.1.1. The result is (a) an object of type ‘referential-category’, and usually (b) one or more context free phrase structure rules (§4).

Defining a category entails specifying the following items. Strictly speaking only the name is required, but useful ones will provide slots for storing information about the category (variables) and information about what rules should be created (realization specifications).

1. Giving it a name.
2. Deciding where it should fit in the taxonomic inheritance lattice of the overall ontology (optional – :specializes).
3. Specifying how individuals that have this category as one of their types will be recorded in the discourse history (optional – :instantiates, see §9.6.).
4. Defining a set of ‘variables’ (§9.???) and their value restrictions (optional – :binds)
5. If individuals (§9.???) with this category are ever created by rule, they have to be ‘interned’ so that subsequent instances will all refer to the same object, in which case the :index field has to be filled out (§9.1.3).
6. Specifying how the we might recognize references to the category within a text, i.e. what rules should be written for it (:realization – §9.5).