26a [Supplementary] Subsumption: Some Formalities

Drew McDermott `drew.mcdermott@yale.edu`

`2015-11-20, 11-30, 2016-10-28` — Yale CS470/570

The purpose of this note is to clarify and simplify some of Shieber's concepts and notations with regard to unification grammars.

Feature structures (henceforth *F-structures*) may be thought of as DAGs, but also as nodes in a DAG. Often the distinction is irrelevant; a node determines a subDAG, by treating it as the root of the subgraph obtained by collecting all the nodes reachable from it. But sometimes the distinction is important, as in the definition of subsumption:

> ... [A] complex feature structure D subsumes a complex feature structure D' [$D \sqsubseteq D'$] if and only if $D(l) \sqsubseteq D'(l)$ for all $l \in dom(D)$ and $D'(p) = D'(q)$ for all paths p and q such that $D(p) = D(q)$. (p. 12)

The first clause, regarding all $l \in dom(D)$, is a local property of D and D', considered as nodes, since it refers only to their children (although the whole paragraph is part of a recursion that makes sense only when you see the definition of subsumption for *atomic* F-structures). In the second clause of this definition, regarding all *paths* through D and D', makes sense only if D and D' are thought of as complete DAGs. Furthermore, there is really no need to repeat it for the child nodes if they should turn out to be complex, because if it holds for the root node of a DAG D, then it holds for all subDAGs of D.

To describe a DAG, we need a way of referring to its nodes. Shieber fails to make the distinction, which makes his diagrams difficult to make precise sense of. For instance, leaf nodes sometimes seem just to be values. But suppose two leaf nodes are supposed to be equal even though their values are not known. There's a notation for that:

...=*n*:{}
...=*n*

At two (or more) different spots the same node must occur, as indicated by the coindexation numbers "=*n*". One of these is marked as a variable: "{}". But once the variable is instantiated, to the value "`plural`," let's suppose, is there a distinction between

...f:=*n*`plural`
...g:=*n*=`plural`

and

```
...f: plural
...g: plural
```

in Shieber's notation? In the previous note, I argued that there was not.

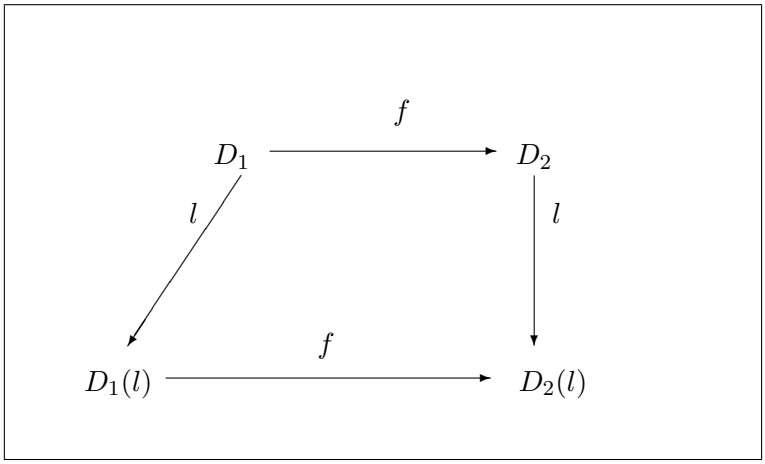
So from now on I propose to use the letter D, with subscripts and and primes, to refer to F-structures considered as *nodes* of a graph. The *F-DAG* the nodes belong to (referred to with the letter G), is a tuple $\langle N, R\rangle$, where N is a set of nodes and $R \in N$ is the root node. Some nodes are *leaves* and have values; if D is a leaf, its value is $v(D)$. (It's often more convenient to have an adjective meaning "leafish," and Shieber's "atomic" will serve this purpose.) The non-leaves are *complex*, and have *child functions* $c(D) : \mathit{labels} \rightarrow N$. But we'll always follow Shieber and write $D(l)$ rather than $c(D)(l)$. We'll also use his notation for *paths* from a node. A path is a sequence of labels $\langle l_1\, l_2\, \ldots,\, l_k\rangle$. I'll use a plus-sign to indicate appending two paths: $\langle l_1\, \ldots, l_j\rangle + \langle l_{j+1}\, \ldots\, l_k\rangle = \langle l_1\, l_2\, \ldots,\, l_k\rangle$. If p is a path and l is a label, $p + l = p + \langle l\rangle$.

$D(p)$ is the node reachable from D by following path p, if all the labels work out; in that case we say $D(p)$ *exists*, or that p is a *well-defined path* from D. $D(\langle\rangle)$ always exists and $= D$. If $D(p)$ exists and $l \in \mathit{dom}(D(p))$, then $D(p + l)$ exists and $D(p + l) = D(p)(l)$.

Shieber's definition of subsumption can be simplified. First, we want subsumption to be a relation between two DAGs, $G_1 = \langle N_1, R_1\rangle$, and $G_2 = \langle N_2, R_2\rangle$. Using the new notation, introduce the term *subsumption homomorphism from G_1 to G_2* to describe a function $f : N_1 \rightarrow N_2$ with the following properties:

1. $f(R_1) = R_2$
2. If $D_1 \in N_1$ is a leaf, then $D_2 = f(D_1)$ is a leaf and $v(D_2) = v(D_1)$.
3. If $D_1 \in N_1$ is complex, then $D_2 = f(D_1)$ is complex, $\mathit{dom}(D_1) \subseteq \mathit{dom}(D_2)$ and $f(D_1(l)) = f(D_1)(l) = D_2(l)$.

The last bit can be graphed as follows:



Then $G_1 \sqsubseteq^\tau G_2$ if and only if there is a subsumption homomorphism from G_1 to G_2. (Bouma 1992 uses the term "homomorphism" to define subsumption.)

We'd like to prove that this defines the same relation as Shieber's definition. We'll write this relation $G_1 \sqsubseteq^\sigma G_2$, and rewrite its definition in terms of DAGs:

1. G_1 *path-subsumes* G_2: For all paths p and q such that $R_1(p)$ and $R_1(q)$ exist and $R_1(p) = R_1(q)$, $R_2(p)$ and $R_2(q)$ exist and $R_2(p) = R_2(q)$.

2. G_1 *recursively subsumes* G_2: $R_1 \sqsubseteq^\sigma_D R_2$

where the relation $\sqsubseteq^\sigma_D$ on F-structures is defined as before:

3. If D_1 and D_2 are atomic, $D_1 \sqsubseteq^\sigma_D D_2$ iff $v(D_1) = v(D_2)$

4. If D_1 and D_2 are complex, $D_1 \sqsubseteq^\sigma_D D_2$ iff $dom(D_1) \subseteq dom(D_2)$ and for all $l \in dom(D_1)$, $D_1(l) \sqsubseteq^\sigma_D D_2(l)$.

5. If one of D_1 and D_2 is atomic and the other is complex, then $D_1 \not\sqsubseteq^\sigma_D D_2$

As before, this part of the definition is completely "local," in the sense that it can be thought of as a simple recursion. It is defined for any two nodes $D_1 \in N_1$ and $D_2 \in N_2$. The paths from the roots to D_1 and D_2 are irrelevant.

In what follows, we'll just say "path to D" when we mean "path from R to D," assuming that the identity of the root R of the F-DAG containing D is understood.

Theorem 1 *For all F-DAGs G_1 and G_2, $G_1 \sqsubseteq^\sigma G_2$ if and only if $G_1 \sqsubseteq^\tau G_2$.*

If you think the theorem is obviously true, you can skip the rest of these notes. On the other hand, phrasing it so it is provable is suprisingly tricky, so perhaps it's not so obvious after all.

Proof: (Only if) Assume $G_1 \sqsubseteq^\sigma G_2$. We'll construct a subsumption homomorphism from $G_1 = \langle N_1, R_1 \rangle$ to $G_2 = \langle N_2, R_2 \rangle$ by defining a series of functions (sets of ordered pairs) $f_0, f_1, \ldots$, such that for some k, f_k is the homomorphism.

Take $f_0 = \{\langle R_1, R_2 \rangle\}$. Then define, for all $i \geq 0$,

$$f_{i+1} = f_i \cup \{\langle D_1, D_2 \rangle : \begin{array}{l} \exists \text{ a pair } \langle D_1', D_2' \rangle \in f_i, \text{ both } D_1' \text{ and } D_2' \text{ complex,} \\ \quad \text{and an } l \in dom(D_1') \text{ such that} \\ \quad D_1 = D_1'(l),\ D_2 = D_2'(l)\} \end{array}$$

We prove by simultaneous induction the following lemma:

Lemma 2 *(a) For all $i \geq 0$, f_i is a function, and (b) if $f_i(D_1) = D_2$ and there is a path p of length i to D_1, then p is a path to D_2.*

Proof of lemma 2: It's obvious for f_0. So assume it's true for all $j \leq i$, with an eye toward proving it's true for f_{i+1}. Consider two ordered pairs $\langle D_1, D_{21} \rangle \in f_{i+1}$ and $\langle D_1, D_{22} \rangle \in f_{i+1}$, at least one of which was not $\in f_i$. There must be two pairs $\langle D_1^a, D_2^a \rangle \in f_i$ and $\langle D_1^b, D_2^b \rangle \in f_i$ (not necessarily distinct), and labels l_1^a, l_1^b (not necessarily distinct) such that $D_1 = D_1^a(l_1^a)$ and $D_1 = D_1^b(l_1^b)$. Because f_i is a function, by induction hypothesis),

$$\begin{aligned} D_2^a &= f_i(D_1^a) \\ D_2^b &= f_i(D_1^b) \end{aligned}$$

So $D_{21} = D_2^a(l_1^a)$, $D_{22} = D_2^b(l_1^b)$. For f_{i+1} to be a function, it must be the case that $D_{21} = D_{22}$.

By construction, there is a path p_1^a to D_1^a of length i or less, and similarly a path p_1^b to D_1^b. By induction hypothesis, p_1^a and p_1^b are also paths to D_2^a and D_2^b, respectively. $D_2^a = R_2(p_1^a), D_2^b = R_2(p_1^b)$. So there are two (not necessarily distinct paths) $p_1^a + l_1^a$ and $p_1^b + l_1^b$ from D_1. $R_1(p_1^a + l_1^a) = R_1(p_1^b + l_1^b)$. See figure 1. By the definition of $\sqsubseteq^\sigma$, clause 1 (remember that?), $R_2(p_1^a + l_1^a) = R_2(p_1^b + l_1^b)$. But $R_2(p_1^a + l_1^a) = D_2^a(l_1^a) = D_{21} = R_2(p_1^b + l_1^b) = D_2^b(l_1^b) = D_{22}$. So f_{i+1} is a function.

The other half of the simultaneous induction is easier. Every path of length $i + 1$ to nodes in N_2 is examined because we've got every path of length i or less in f_i. Whenever the ordered pair $\langle D_1, D_2 \rangle$ is in $f_{i+1} \setminus f_i$, it's always by adding the same label l to the paths to the parents of D_1 and D_2 in f_i. QED(lemma 2)

Under construction
(Or look at the previous diagram and cross your eyes.)

Figure 1: Complicated diagram

Using this lemma, we can focus on f_*, defined as f_k where k is the lowest subscript for which $f_k = f_{k+1}$. (There must be such a k, because F-DAGs are finite.) It remains to prove that f_* is a subsumption homomorphism. In every pair $\langle D_1, D_2 \rangle \in f_*$ either both D_1 and D_2 are atomic, or both are complex. (This follows easily from part (b) of Lemma 2.) In the complex case, by construction all the children of D_1 and D_2 satisfy the homomorphism property, $f_*(D_1(l)) = D_2(l)$. The atomic case arises only for nodes at the ends of terminal paths. Here is where we use the fact that G_1 recursively subsumes G_2, i.e., that $R_1 \sqsubseteq^\sigma_D R_2$. Lemma 2 shows that the sequence of recursive tests that define whether $R_1 \sqsubseteq^\sigma_D R_2$ tracks the paths kept in synch by f_*. The truth of $R_1 \sqsubseteq^\sigma_D R_2$ ultimately depends on the truth of $D_1 \sqsubseteq^\sigma_D f_*(D_1)$, where D_1 is a leaf. If $f_*(D_1) = D_2$, and D_1 is a leaf, $D_1 \sqsubseteq^\sigma_D D_2$ can be true true only if D_2 is a leaf, and $v(D_1) = v(D_2)$.

Hence f_* is the subsumption homomorphism we seek. QED("Only If")

Now for the other half of Theorem 1, the part traditionally called "(If)." Assume $G_1 \sqsubseteq^\tau G_2$, so that there is a subsumption homomorphism f from G_1 to G_2. We need to show that G_1 path-subsumes and recursively subsumes G_2.

The proof that G_1 path-subsumes G_2 depends on another lemma:

Lemma 3 *For every path p, if $R_1(p)$ exists then $R_2(p)$ exists and $f(R_1(p)) = R_2(p)$.*

Proof of lemma: By induction on the path length. It's obvious for paths of length 0. Assume it's true for a path of length i, and let $p = p' + l$ be a path of length $i + 1$ such that $R_1(p)$ exists. Then $l \in dom(f(R_1(p')))$, and $f(R_1(p)) = f(R_1(p'))(l)$, which $= R_2(p')(l)$, if $l \in dom(R_2(p'))$. But of course the fact that f is a subsumption homomorphism means that $dom(R_1(p')) \subseteq dom(R_2(p'))$, so we're done. QED(lemma 3)

This lemma may be said to be the whole point of subsumption homomorphisms. To prove path-subsumption, we need to show that if there are two paths p^a and p^b such that $R_1(p^a) = R_1(p^b)$ then $R_2(p^a) = R_2(p^b)$ But this follows immediately from the lemma:

$$R_2(p^a) \quad = \quad f(R_1(p^a))$$

$$
\begin{aligned}
&= f(R_1(p^b)) \\
&= R_2(p^b)
\end{aligned}
$$

We still have to prove that G_1 recursively subsumes G_2, i.e., that $R_1 \sqsubseteq^{\sigma}_{D} R_2$.

We do it by induction on the height of G_1. The *height* of G_1 is the height of R_1. The height of a node is defined as follows:

1. If D is a leaf, its height is 0.
2. If D is complex, its height is 1 + the maximum height of any of its children.

If R_1 has height 0, it's a leaf. So $f(R_1) = R_2$ is a leaf and $v(R_1) = v(R_2)$. So $R_1 \sqsubseteq^{\sigma}_{D} R_2$.

Now assume (induction-hypothesis alert) that for every F-DAG $G'_1 = \langle N'_1, R'_1 \rangle$ of height $\leq i$ that comes equipped with a subsumption homomorphism f' to another F-DAG, $G'_2 = \langle N'_2, R'_2 \rangle$, it is the case that $R'_1 \sqsubseteq^{\sigma}_{D} R'_2$. Our goal is to show that for an F-DAG $G_1 = \langle N_1, R_1 \rangle$ of height $i + 1$, again, accompanied by a subsumption homomorphism f to $G_2 = \langle N_2, R_2 \rangle$, $R_1 \sqsubseteq^{\sigma}_{D} R_2$. Since R_1 is complex, the existence of f means that $dom(R_1) \subseteq dom(R_2)$.

But there is a slight hitch. Given the way we define F-DAGs, G_1 does not have components that are themselves F-DAGs. But we can pretend that it does by defining $N(D)$ as all the nodes reachable from D (including D itself), so that $\langle N(D), D \rangle$ is the F-DAG rooted at D; call this $G(D)$. We can also define $f \upharpoonright D$ as $\{\langle D_1, f(D_1) \rangle : D_1 \in N(D)\}$. So if C^i_1 is a child of R_1, $C^i_1 = R_1(l^i)$, $G(C^i_1)$ is an F-DAG of height $\leq i$, $f \upharpoonright C^i_1$ is a subsumption homomorphism from $G(C^i_1)$ to $G(f(C^i_1)) = G(R_2(l^i))$, and the antecedent of the induction hypothesis is satisfied. Hence for all $l^i \in dom(R_1)$, $R_1(l^i) \sqsubseteq^{\sigma}_{D} R_2(l^)$. By clause 4 of the definition of $\sqsubseteq^{\sigma}_{D}$, $R_1 \sqsubseteq^{\sigma}_{D} R_2$. QED("If")
QED(Theorem 1)