Uniformity of Representation

We can use three ways of saying things:

> The wordgroup: Helicopter rotor
>
> Prepositions: The rotor of the helicopter
>
> Verbs: The helicopter has a rotor

If we use a common form of representation, these things will all look the same. That means we need to unpick the structure of the wordgroup, and figure out the different meanings of a preposition – "of" has about 15.

We can use identity to make things easier:

> Object-attribute ⇔ attribute of object

That is, wordgroups are equated with their prepositional equivalents, so a direct relational structure of the wordgroup is not necessary.

It slowly dawned that prepositions are mostly remnants of collocated verbs, as

> Anesthesia that is required for surgery that is to be performed on body part by surgeon
>
> Anesthesia for surgery on body part

If we put back the relations, we would have a consistent structure, without the need to figure out what the preposition means when we encounter it in the semantic structure.

Both wordgroups and prepositional phrases can be hierarchical, so "body part surgery" and "surgery on body part" cover thousands of different body parts, and thousands of different types of surgery, and allow specific examples to be built. This implies constant automatic reorganisation – "subtarsal fusion" fits under "body part surgery", but if "ankle surgery" (or "surgery on ankle") is added, then "subtarsal fusion" needs to move under that.

Some of the meanings of "of"

RelationObject	analysis of data	unknown person analyses data
RelationSubject	the excitement of the visit	visit excites unknown person
ToBeAttributeOf	Attribute	temperature of body
Reverse Have	disease of lung	lung has disease
ToBePerformedOn	arthroplasty of knee	unknown person performs…
Quantity of thing	six of them	thing.count = 6
RelationSubjectRelationObject	owner of boat	owner owns boat

RelationObjectRelationSubject employee of company company employs employee

Attribute or component is only checked for the initial object, but a chain will be followed, if the start is found.

“owner” needs to be paired with “ownable thing”, so we are not confused by

> “He accused the owner of fraud”

And still get

> “He accused the owner of the boat of fraud”

The advantage of uniformity of representation is that it greatly simplifies understanding the knowledge structure, and navigating through it.

The health version has

40,000 words in the dictionary

20,000 wordgroups (concepts), ranging from “cough medication” to “multi-family group adaptive behaviour treatment guidance”.

10,000 relations, of which 3,000 are collocated verbs (several meanings of take off, look out for, cut up) – about half the nouns associated with collocated verbs (several thousand) can use the verb collocation modelling for handling prepositions, so they don’t need their own prepositional phrases.

5,000 prepositional phrases – “surgery on body part” – many are wordgroup equivalents

Overall, Health uses less than a million network elements – less than 2 million after reading the medical codes.

We were getting tired of adding doctor’s names – Brown syndrome, Sofield procedure, de Quervain’s disease, so we are starting to automate the modelling of unknown words, using Oxford, Medline, Wikipedia. There has never been a better time to do this in terms of available online resources, and we think we have reached the critical mass needed for automatic knowledge extension.

Health has taken four years – about 70 wordgroups a day by hand. The next domain will be mostly done using automation, starting with searching for unknown words, wordgroups and prepositional phrases in domain text. Health is very broad and yet highly detailed, so no other domain is likely to match it.