Tuning Document-Based Hierarchies with Generative Principles

Piek Vossen
Irion Technologies
Bagijnhof 80
2601-CV Delft, The Netherlands,
Piek.Vossen@irion.nl

Abstract

In this paper I describe how a more-consistent and clearer hierarchy can be extracted for specific document collections. First, a hierarchy is automatically extracted from the terms in the documents. This hierarchy is extended with concepts and classifications from WordNet. I explain how this hierarchy can be improved using Generative lexicon principles in combination with the information on the term usage that is extracted from the documents.

Introduction

A hierarchy or tree structure is a powerful way of giving access to large quantities of unstructured information. A user can browse the hierarchy to the concepts in which he is interested and, from there, directly access the documents in which these concepts are discussed, possibly jumping to their occurrences in the documents. Compared to query-retrieval, the tree provides you with a feeling of what can be found or what is present in a document collection. You can directly see the conceptual neighbors of a concept and how specific the classification is. This is especially useful if users do not exactly know what information is present or how the information is structured.

In addition, a hierarchy can be used as the starting point for developing a domain-specific ontology. These ontologies are needed to develop even more sophisticated information systems, intelligent dialogue interfaces or information extraction and data mining technology.

Whereas it is relatively easy to build a search-engine by indexing the documents, it is however lesser clear how to build a classification tree that is useful for these purposes. A general-purpose hierarchy, such as WordNet (Fellbaum 1998), represents a fixed hierarchy and structure. It contains many concepts and distinctions that are not relevant for certain domain-specific document collections. It not only imposes ambiguity of terms but also too much complexity for the intended purpose. Another drawback of using a fixed general-purpose hierarchy is that it only represents a single perspective of classifying the document terms. For building document-specific and customer-specific applications it is necessary to build hierarchies that incorporate the semantic distinctions made in the document and no more than those.

In this paper, I describe how this can be done by augmenting the WordNet hierarchy with a qualia structure and revising the classifications using Generative Lexicon principles (Pustejovsky 1995) and term frequency information. The Generative qualia classification gives maximum flexibility to build a hierarchy from different perspectives. Following Guarino (1998), the different status of taxonomic classifications and their document frequency is then exploited to build a hierarchy that is consistent, relevant and clearly structured.

The structure of this paper is as follows. In section 1, I shortly describe how document-based hierarchies are built. Section 2 discusses some of the problems for these hierarchies. Section 3 explains how the initial hierarchy can be revised.

2. Extracting hierarchies from documents.

The extraction of the initial hierarchy is done by the following procedure:

1. Extract the most significant NPs from shallow-parsed text.
2. Extract all salient and lexicalised multiword sequences from the NPs.
3. Decompose the multiword sequences into head-modifier structures.
4. Fill the database with concepts from WordNet.
5. Build a hierarchy that combines the decomposition information with the concept hierarchy from WordNet.
6. Trim the meanings of the document terms to the relevant concepts only.

Details about the procedure can be found in Vossen (2001fc). Here I will only give a short description. All extracted NPs and their head-modifier structures are stored in a central database as so-called topics. The head-modifier relations are stored as topic-to-topic relations. For selecting appropriate terms a salience measure is used, based on the document frequency and lexicality (Justeson and Katz 1995). A similar measure is used to determine the chunking of the head-modifier relations. A preference is given to lexicalized and salient parts. Below is an example of the chunking of extracted multiword terms related to *technology*:

```
technology
  printing technology
     digital printing technology
     smart printing technology
  inkjet technology
     next generation inkjet technology
     inkjet technology through third parties
     world leading inkjet technology
     thermal inkjet technology
     hp thermal inkjet technology
  color layering technology
     ii color layering technology
     photoret iii color layering technology
     color technology
```

Instead of a hierarchy where all multiwords are directly related to *technology*, we create 3 levels by following the salient topic-to-topic head relations in the database.

The above tree is completely based on the decomposition of multiwords. The tree will consist of as many tops as there are single words in the term database. This can be a few thousand for the collections of documents that we process.

There are several reasons why we would like to augment these compositional trees with a semantic network as WordNet:

1. WordNet synonyms can be used to cluster or merge nodes and thus branches in this tree;
2. WordNet can be used to reduce the number of tops by adding classifications of tops and intermediate levels;

We import WordNet synsets as separate concept records into the database together with their concept-to-concept relations (relations between synsets). All the imported concepts are linked to the topics if there is a match between the topic variants and the concept variants. Topics will get a list of concept references and concepts a list of topic references. There will also be concepts without topic references and topics without concept references.

In addition to the previous tree that was built from topic-to-topic head relations, we can now also build trees based on the concept-to-concept hyperonym relations from Wordnet. However, it is also possible to combine a tree of topics with a tree of concepts by including both the topic-to-topic and concept-to-concept relations. Likewise, we can extend the above *technology* tree with the concept relations from WordNet, or, vice versa, extend a WordNet hierarchy with new terms decomposed via topic-to-topic relations:

```
psychological feature 1
  cognition 1
    cognitive content 1
      knowledge base 1
        branch of knowledge 1
          technology 2
            printing technology
            inkjet technology...etc..
act 1
  activity 1
    employment 2
      application 3
        technology 1
          printing technology
          inkjet technology...etc...
```

There are two different classifications for technology because there are two different meanings in WordNet. By simply merging the tree of topic relations with the tree of concept relations, we will thus duplicate the topic subtrees at every meaning of every concept.

There will also be a reduction of branches in the tree because of the collapse of synonyms. Since *engineering* belongs to the same synsets as *technology*, all topics related to *engineering* will be linked to the same concept as the topics linked to *technology*.

The term *technology* only has two meanings in WordNet, but others have many more. Especially if polysemy occurs at several levels, this leads to an explosion of terms in the hierarchy. We therefore trim the trees by limiting the concepts to the particular context. For disambiguating the topics, we make use of their frequency information and the glosses in Wordnet. Following Mihalcea and Moldovan (1999), the glosses in Wordnet are used as a context definition. However, instead of comparing the words in the glosses with the context in the text, we weight the words in the gloss using their frequency in the documents set, compared to the frequency in all the glosses:

$$\rho(w) = \frac{df(w)+ef(w)}{gf(w)}$$

The probability **ρ** of a content word **w** in the gloss is obtained by cumulating its document frequency **df** with the element frequency **ef** and dividing the sum by the frequency of this word in all the glosses of Wordnet: **gf**. A word has a high probability, if it occurs frequently in the document set compared to its overall frequency in the glosses.

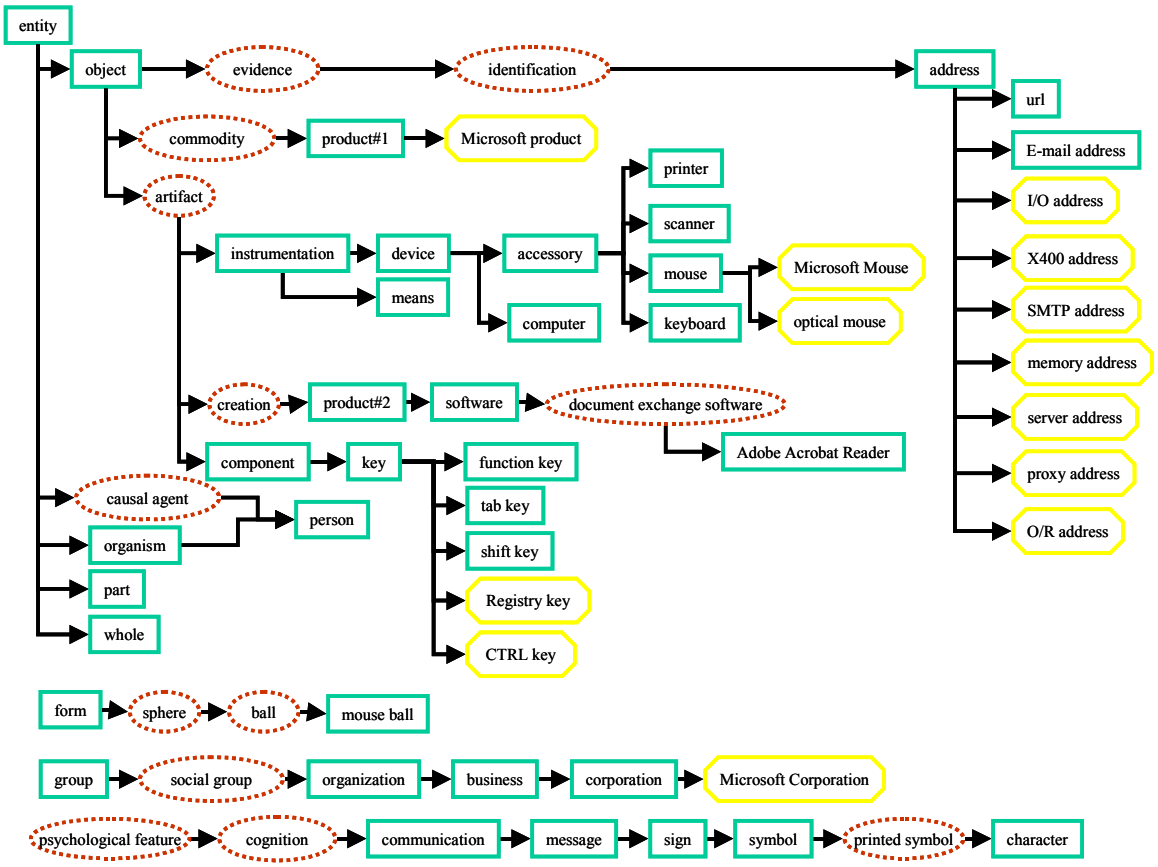


Figure 1: Hierarchy that combines new terms with WordNet classifications.

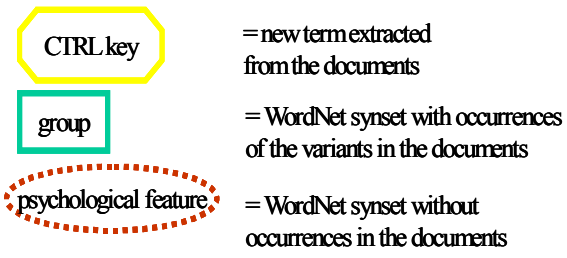
The probability of a concept (**C**) is then based on the sum of the probability of the content words ($\mathbf{w_i}$), divided by the total number of content words (**N**) in that gloss:

$$\rho(C) = \frac{\sum_{w_i}^{N} \frac{df(w_i)+ef(w_i)}{gf(w_i)}}{N}$$

For each topic, we then removed all concepts with probability less than 75% of the maximum probability of the topic. See Vossen (2001fc) for an evaluation of the trimming.

When we build a combined tree of topic-to-topic and concept-to-concept relations, we get a small chunk of WordNet that is extended with newly extracted terms. Such a hierarchy can be seen as a micro-wordnet for a specific domain or document collection. Figure-1 shows a fragment of the upper hierarchy that you will get when the term-hierarchy is combined with a trimmed concept-hierarchy from WordNet. The nodes have been selected to illustrate some phenomena that I will discuss below. The discussion is also limited to concrete concepts.

Different shapes and colors have been used to differentiate the origin and status of the nodes. Yellow octagons represent new terms that do not occur in WordNet and that are extracted from the documents. The green squares represent synsets from WordNet that have one or more variants occurring in the documents. Finally, the red dotted ovals are synsets that do not occur in the documents but that are generated as the hyperonyms of other synsets that do have occurrences:



Usually, the top levels are red dotted ovals, the middle levels are green squares and lower levels are yellow octagons. Occasionally, red ovals and green squares are mixed. Yellow octagons can only be linked to green squares.

We see four unique beginners from WordNet at the left side of Figure-1: *entity*, *form*, *group* and *psychological feature*. Most of nodes are green squares. Here and there we see an oval red hyperonym. At the right side we see a mixture of green squares or yellow octagons. The terms have been extracted from the Microsoft Support side. When applied to large collections of technical support documents in the Information-Technology branch (30,000 documents), we typically get distributions such as:

- 20,000 yellow octagons
- 4000 green squares
- 500 red ovals

In the next sections, we will mainly discuss this upper-level of the hierarchy. All the middle and lower relations can usually be taken over from WordNet. Before we evaluate the top-level classification in the next section, I want to make two comments. Firstly, WordNet was not designed for the purposes that we are using it for. Therefore, the comments made here should not be seen as a critique on WordNet. Furthermore, the phenomena we see here are common to any Sense Enumerating Lexicon (SEL, Pustejovsky 1995). Unfortunately, WordNet is the most easily verifiable example. Secondly, I used an edited version of WordNet1.5 that has been extended and changed by lexicographers in the company to adapt it specially to the Information Technology domain.

2. Evaluating the top-level classifications

As said in the introduction, the hierarchy is used for two purposes:

1. To develop an ad-hoc ontology
2. To provide users with a classification access to documents.

An ad-hoc ontology covers the distinctions made in the documents and no more than that. Semantic distinctions that are relevant should be applied to all concepts that carry this distinction

and no other concepts than those. The ontology thus has to be:

- Correct: all expressed implications are valid.
- Consistent: implications are expressed systematically.
- Comprehensive: all possible implications are expressed.
- Efficient: distinctions should be introduced only once.

The first requirement is mostly met. It hardly occurs that a wrong implication is expressed. We find many more violations of the second requirement. Consider for example, printed and physical representations of language and communication at the bottom of Figure-1, which are linked to communication as a cognitive phenomenon. First of all, graphical representation is a physical phenomenon resulting from cognitive and physical behavior. Secondly, we miss here the implication that symbols are physical objects as well. Further down the hierarchy of *sign* and *symbol,* we find many physical objects such as *books* and *documents*.

The second example is *mouse ball,* which has been linked to a sense of *ball* that is not related to *artifact* and *object*. The object sense of *ball* is however limited to *ball used in games*. Here we see a situation where there is no general abstract meaning of *ball* that expresses both its *shape* and *object* properties, regardless of its purpose. The lexicographer apparently made a choice for the most general meaning but a *mouse ball* **has** a *ball-shape* and it is **not** a subtype of it.

Similar things can be said for other concepts, e.g. *address* is not linked to *symbol, artifact* and *object.* Strictly speaking, we should also systematically apply the 3-fold differentiation, *part*, *group* and *whole* to all more specific concepts. Any more specific concept is either an independent *whole*, a *part* of some other entity or a *group* of other entities. We see that for many concepts this is not expressed. An obvious case is again *mouse ball*, which a component of a *mouse* and thus should be classified accordingly as a *component.*

These implications have been missed because most concepts are classified with a single hyperonym. This is typical for a SEL resource. Rather than describing concepts as complex constructs that can be defined by multiple qualia, only one of them is selected, while different senses reflect separate dimensions of classification: functional, constitutional, or agentive hyperonyms. We already saw this for the different senses of *ball*. It would have been sufficient if there is just one sense of *ball* that is defined as a *spheric object*. The origin (artifact or natural) and the purpose can then be specialized in context. Subtypes of *ball*, such as *football* and *volleyball* have an additional functional and agentive classification. In the case of *mouse ball*, there will then be additional functional, agentive and constitutional classifications:

mouse ball $\Rightarrow$ *ball*
$\Rightarrow$ *device*
$\Rightarrow$ *artifact*
$\Rightarrow$ *component* in *mouse*

Another strategy to express multiple classifications in a SEL resource is the introduction of restricted senses that combine two or more classifications. In Figure-1, we see for example that a separate sense of *component* is introduced below *artifact* to capture the fact that some hyponyms are both *components* and *artifacts*. This has several consequences:

1. *Component* is not linked to *part*, so we cannot get at all parts via *part*.
2. The semantics of *part*-hood has to be defined at two places in the hierarchy: violating the third requirement.
3. The hierarchy becomes deeper and more complex.

The same holds for the specialized sense of *accessory*, which is introduced below *device*.

Note that there is one example of multiple hyponymy: *person* is both classified as *causal agent* and *organism*. On the other hand, *causal agents* are not restricted to concrete entities. *Events, phenomena, psychological states* and *ideas* can be *causal agents* as well. Here we see

another case of too limited usage. The semantics of *cause* is not maximized, violating requirement 3. The same can be said for *artifact*. It is limited to *object* but there can be *artifact substances* as well.

If we look at the hierarchy, from a practical point of view, as a classification interface to documents, there are some other additional requirements:

- The hierarchy should contain the relevant distinctions only: users should not have to look at many nodes that are not related to the topic.
- The hierarchy should not be too deep: users should be at the concepts in just a few steps.
- The hierarchy should not be too complex: the graphical display should be a tree and not a tangled lattice.
- The hierarchy has to be intuitive: users should find nodes at places where they expect them.

The first requirement is too a large extent met by the followed procedure. Hyponyms of categories are only represented if there is evidence in the documents. This means that the subtrees below *object*, *artifact* and *organism* are limited to concepts that occur in the documents, whereas they otherwise contain hundreds of specific concepts.

Still, we see that some of red oval hyperonyms do not seem to add interesting classifications: *social group, psychological feature, creation, evidence*. Even though their semantic implication may be valid, it still is to be seen if the distinction is relevant for the ontology and the way it is used (in for example information extraction tasks).

Secondly, the depth and inconsistency of the hierarchy, discussed above, make it complex and unintuitive for users. Users that browse this hierarchy will not find all *devices*, *symbols*, *artifacts* or *objects* below the nodes in the hierarchy.

On the other hand, the fact that WordNet is to a large extent a tree can also be seen as a positive feature. A tree is easier to grasp than a tangled lattice. There is only one route to a concept, although you can only find the concept if you know the path.

Finally, there is also an example of an unused classification that may have been interesting from a user-perspective: *Microsoft products*. A hierarchy that is extracted from a Microsoft support site will contain many terms that could be classified as *Microsoft products* but there is no way in which this can be anticipated in a generic resource. Nevertheless, it is possible to extract this classification by imposing *Microsoft product* as an additional hyperonym on certain classes of terms that have *Microsoft* as its modifier: *Microsoft Mouse*.

To summarize, we have seen the following:

- Mostly, a single hyperonym is given.
- Classifications have not been applied systematically.
- Specialized meanings are created to add multiple classifications in sub levels.
- Relatively deep and complex tree structures are used.
- Still, some irrelevant classifications are generated that are not used in the documents.

In the next sections, I will discuss how we try to improve this hierarchy using a top-ontology based on the qualia model in the Generative Lexicon (Pustejovsky 1995).

3. Restructuring the top-level classifications

To restructure the above top-hierarchy, we first create a more consistent and richer upper level classification. Secondly, we make use of the different status of the classification to limit and simplify the visualization of the classification for users. We more or less follow a 3-stage procedure:

1. Add a qualia classification to the WordNet top-hierarchy and revise the hyperonym relations.
2. Allow the extraction of additional user-defined classifications.

3. Limit the visualization of the hierarchy according to rigidity, dependency and relevance of the classification.

We first impose a qualia classification to the top-level synsets of WordNet, as is done with the top-ontology in EuroWordNet (Vossen 1998). This has the effect that parallel qualia classifications are systematically imposed on the hierarchy. We can then restructure some of the above classifications. Next, it is possible for users to add additional classifications that can be extracted from the documents or are of special interest. Finally, we generate a limited hierarchical structure that selects from the potential classifications certain categories, differentiated for the status of the relations and their relevance to the documents.

3.1. Qualia structure as a top-ontology

In EuroWordNet, a top-ontology was developed based on the qualia in the Generative approach. The ontology was specifically developed to provide a common framework for important concepts or synsets in wordnets for different languages: English, Spanish, Italian, Dutch, German, French, Czech and Estonian. Importance of concepts was based on a mixture of criteria, such as: level in the wordnet hierarchy and the number of children. The top-ontology thus has been empirically validated to cover at least the distinctions that play an important role in these wordnets.

For the lexical semantics working group of Eagles (Sanfilippo et al. 1999) and for the Ansii Committee for Ontology Standards (Hovy 1998), we further elaborated the set of basic concepts and the qualia ontology. The result is given in Figure-2 below. The ontological notions are all preceded by an omega: Ω, to differentiate them from synsets in WordNet.

As in the Generative Lexicon, a distinction is made in 4 different qualia:

- Role (Telic)
- Form (Formal)
- Structure (Constituency)
- Origin (Agentive)

The qualia express complementary aspects of the meaning of words. Words can thus have specific values for each of the 4 qualia. The main qualia have been further sub-divided. In the case of Form and Origin, these subdivisions are strictly disjoint. Role and Structure divisions are orthogonal.

There are then a few differences with respect to the EuroWordNet top-ontology. In EuroWordNet, the qualia Role, Form and Origin were restricted to concrete Things. This turned out to be too restrictive. Many of these semantic notions can also be applied to Situations and Concepts, e.g.:

natural phenomena => Ω-Origin-Natural
force => Ω-Role-Force
plan, method => Ω-Role-Usage

Notions such as *part* and *group*, or *natural* and *artificial* can be applied to many more concepts than just concrete objects.

Furthermore, we can drastically reduce the ambiguity of some of the synsets if we generalize their meaning to a single qualia and apply their semantics orthogonally. Whereas we find different meanings for these words in SEL resources (*part of event, part of substance, part of an object*, etc.), it is now thus sufficient to distinguish a single meaning and derive the specific meanings compositionally or from the context.

All the top-level synsets in WordNet are then linked to their respective quale. The classification generalizes their semantics as much as possible. Specialized meanings are avoided, where possible. More specific synsets are linked to all the top-level synsets, so that all cases of:

$synset_i \Rightarrow q_k \,\&\, q_l$
$synset_j \Rightarrow q_k$

are replaced by:

$synset_i \Rightarrow synset_j \,\&\, q_l$
$synset_j \Rightarrow q_k$

When applied to all upper level synsets in WordNet, we get a gamut of synsets that are only characterized by a single quale and can be applied to any entity (abstract, concrete, event or thing). As explained in Vossen (1995) and Wilks (1996), these upper levels concepts are often

For contrast, the classifications we have discussed so far are individual-level nominals that carry identifying properties of entities to which they refer. Note that some of these still apply in a orthogonal way or may even predicate role-relations. A similar distinction is made by

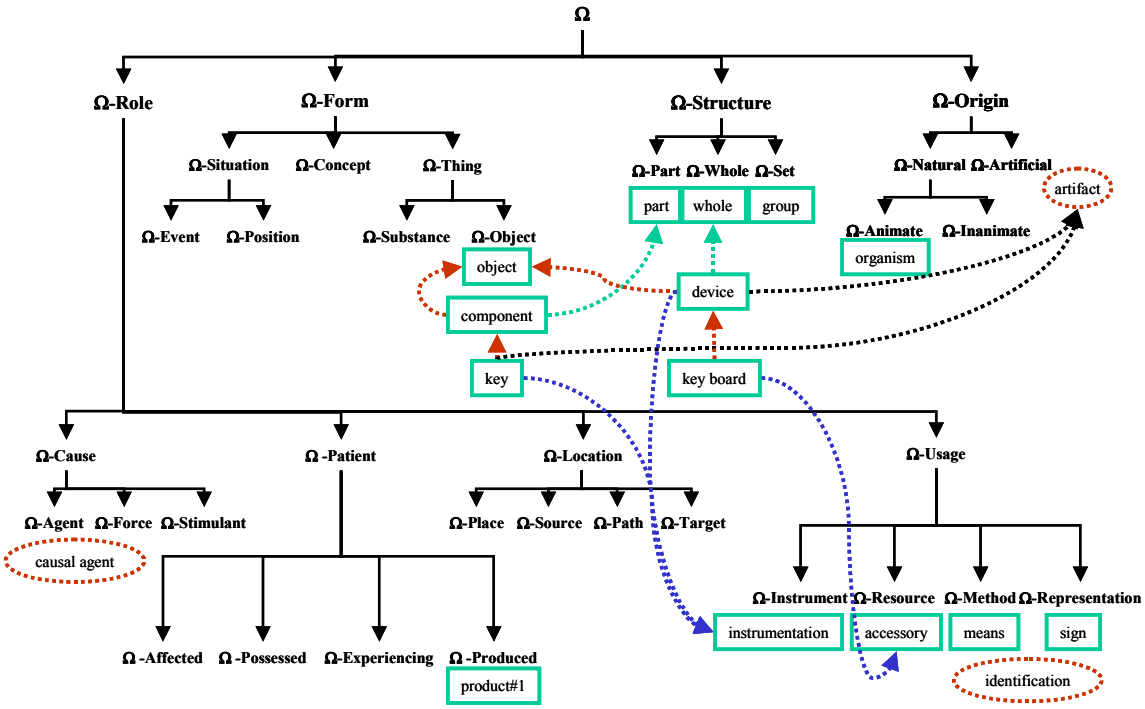


Figure 2: Enhanced EuroWordNet Top-Ontology.

defined in dictionary definitions with void heads: *anything or something that*. We typically see that a selection of these synsets is also used to define more specific concepts but most of these have no specific hyponyms. Some examples given in Vossen (1995) are:

> *buzzer. stiffener, annoyance, attraction, discouragement, threat, target, winner, loser.*

In Figure-1, *commodity-product* represents a similar case of a high-level abstract concept without any or many hyponyms. These nouns can be seen as stage-level or role-defining nominals (Carlson 1977, Pustejovsky 1995). Stage-level nominals express temporal characteristics of entities but no defining characteristics.

Vossen and Bloksma (1998) in terms of strictly disjoint categories and circumstantial categories and by Guarino (1998) between roles and types.

When we apply the above re-organization to the hierarchy in Figure-1, some intermediate levels, such as *artifact*, *accessory*, *component* are taken out of the *object* subtree and are applied separately to the relevant concepts: as is done for *key* (artifact) and *keyboard* (accessory). The effect will be that we get a more complex lattice with multiple hyperonym classification but less levels. We can thus express the same semantics as in Figure-1, using lesser levels in the hierarchy.

Furthermore, we will maximize the meanings of synsets by placing them at the most generic level: *part, group, artifact*, etc. Likewise, there

is no need to differentiate specialized meanings for *artifact* (limited to objects), *component* (limited to artifact) or *accessory* (limited to devices).

The top-node synsets typically only express one quale. As we can see in Figure-2, most of the top-level synsets of Figure-1 can be matched with ontological notions in the qualia ontology. Below Ω-Origin, we find *organism* and *artifact*, below Ω-Structure *part*, *whole* and *group*, below Ω-Thing *object*, and various of the Ω-Role: *causal agent*, *product*, *instrumentation*, *accessory*, *means* and *sign*. The close match is not surprisingly since the ontology was based on the WordNet top-levels.

Below these, I have inserted some more specific synsets from Figure-1, to demonstrate how combinations of qualia are expressed with multiple hyponymy:

device	⇒ whole
	⇒ object
	⇒ artifact
	⇒ instrumentation
component	⇒ part
	⇒ object
key	⇒ component
	⇒ artifact
	⇒ instrumentation
key board	⇒ device
	⇒ accessory

3.2. Selecting classifications for representation

Now we have a very rich classification of concepts that can be applied to the terms extracted from the documents. We gained consistency and completeness and the number of levels in the hierarchies will be reduced because we generalized certain meanings. The drawback of this measure is however that we still have a complex hierarchy of many multiple hyponymy relations. This hierarchy is not very useful for browsing a classification either. More specifically, the hierarchy does not differentiate between the status of the different classifications. Some classifications are strictly disjoint (Form and Origin), whereas others are less strictly separated.

Guarino (1998) even wants to go a step further. by limiting all hierarchical Type relations to disjoint properties or taxons that carry identity criteria. He tries to formally define the differences in terms of 4 notions:

- Identity: taxons that carry an IC;
- Rigidity: if P is true in one possible world it is true in all possible worlds. *Person* and *location* are rigid, *student* and *tall* are not.
- Anti-Rigid: for each x, P(x) is true in one possible world, and false in a different possible world. Examples: *student* and *tall*.
- Dependence: a property P is dependent if, necessarily, whenever P(x) holds, the property Q(y) holds, with $x \neq y$ and $P \neq Q$. Examples: *father* and *part.*

A Type is a property that is rigid and carries an IC. According to Guarino (1998), Types play the most important role in a taxonomy. A taxonomy of Types is always a disjunct tree. A role is a property that is anti-rigid and is always dependent. No explicit disjointness assumption is made for Roles, as they tend to generate tangled hierarchies. Guarino (1998) states that Roles have limited organizational relevance.

The qualia-lattice in Figure-2 does not formally make a distinction between the four qualia. However, that is a matter of definition. Guarino also defines Role and Structure dependencies as separate relations. For our purposes it is sufficient to know that certain properties are disjoint and rigid, whereas others are stage-level properties and/or dependent. We can then use Guarino's principles to present the hierarchy to a user in a consistent and natural way.

We have formulated the following principles for presenting the semantic structures:

1. Only present categories that occur in the documents.
2. Only disjoint and rigid relations are presented in the subtype hierarchy: i.e. those taxons that link to the disjoint Ω-Form or Ω-Origin subtrees.

3. Only non-dependent concepts are listed in the hierarchy of Types.
4. Dependent Ω-Structure relations are expressed towards the entities on which they are dependent.
5. Ω-Role sub-hierarchies are not expressed.

Even though, WordNet provides a rich fund of both stage-level and individual-level nominals (high-level and abstract), the derived hierarchies will only give those upper-level synsets that also occur (frequently) in the documents or are hyperonyms of others that do. Applying these principles to the enhanced hierarchy of Figure-1 will then result in an initial tree that is fairly simple and straight forward. This is shown in the framed area in Figure-3 below. Only the non-dashed lines are shown in the tree. The hierarchy thus has 3 tops:

organism	=> person
object	=> device
	=> software
	=> symbol
organization	=> business

We see that red oval categories are neglected, as well as categories that only express functional or structural dependencies. Structural relations can thus only be accessed via the structural dependency relation that is expressed with the independent disjoint Types. For example, *mouse ball* has as the hyperonyms *ball*, *artifact* and *component*, where the former two would link it to the disjoint hierarchy. However, since it is a dependent entity, we will still not directly list it in the tree of Types. A user can access the concept only via the concept on which it is dependent: namely *mouse*.

There are then two types of hidden categories. Other available categories that are used in the documents, such as:

Ω-Role:	accessory, means, sign, product, instrumentation
Ω-Structure:	component, part, whole, group

Further categories that do not occur in the documents, which are:

Ω-Origin:	artifact
Ω-Role:	causal agent, identification

The hidden links are here represented with dashed lines and dot endings. The user that is structuring the hierarchy can still activate any of these categories and apply it to the subtypes. Individual-level categories, which are expressed but not shown, can automatically be applied to their subtypes: e.g. *sign, accessory, instrumentation, artifact*. Stage-level or circumstantial categories have to be assigned manually: e.g. *product*.

In practice, the user gets an initial proposal for a top-hierarchy as represented in the framed area. He can then modify and activate the categories at will, and design a private top-level classification. For the suggestion of possible categories, he can look at the frequency of these terms in the documents, or even directly go to the occurrences of the terms in the documents to verify the usage. Note that it is thus possible to neglect red oval taxons but also to limit taxons to the most frequent green squares.

Once the upper-levels are designed, a tree fusion function will then take the generated tree and merge it with the private top-level tree. The fusion program works as follows. It will traverse the source tree bottom up or from right-to-left. Whenever it finds a matching node, a so-called interface node, in the target tree, it will cut out the sub-tree from the source tree and place it below the matching node. The fused tree is therefore always compatible with the classification of the target tree. If there is no match, it will go to the next node. The hyperonym relations in the source tree are thus used to get at a level that matches an interface node in the target tree. This also means that source trees are fused regardless of how deep and specific the terminology is. The interface nodes can be specified at any desired level of abstraction.

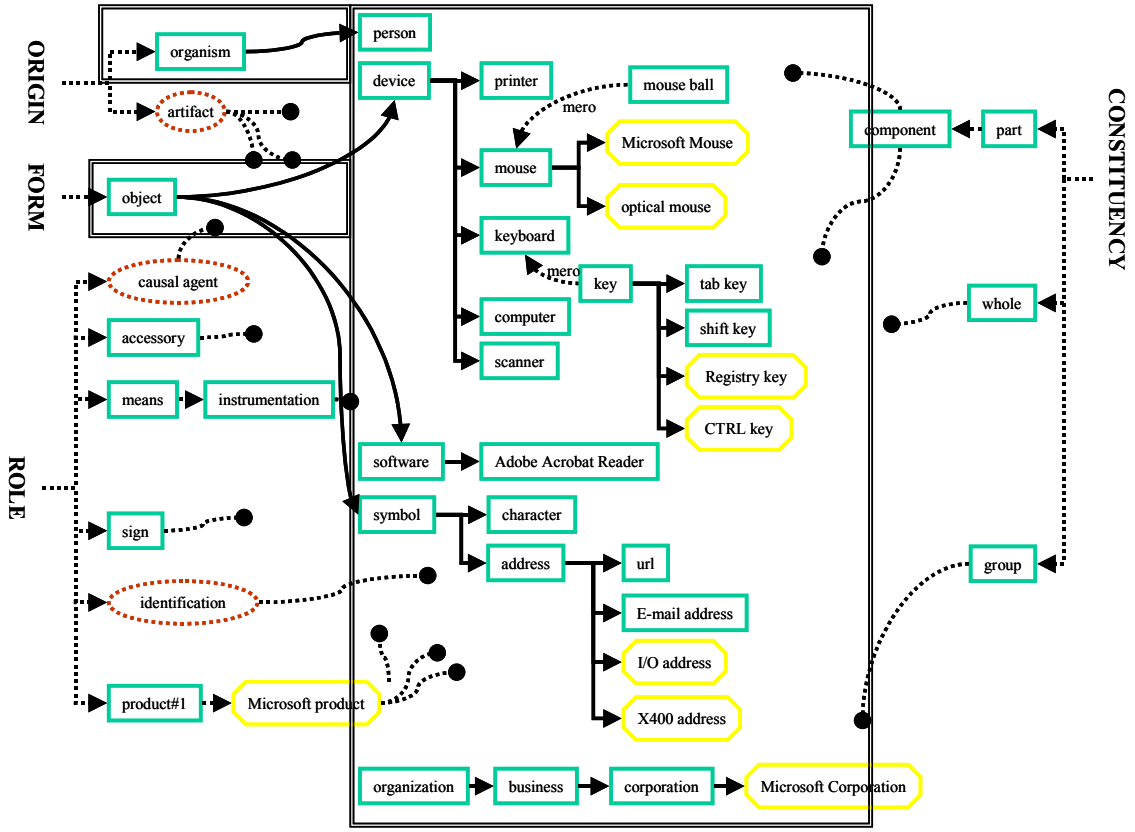


Figure 3: Revised and minimalized top-level hierarchy

The final ontology will thus present a hierarchical structuring of the terms and concepts that occur in the documents that meets the following requirements:

1. It will be correct, consistent and complete due to the systematic qualia classification.
2. It includes a minimal number of levels.
3. It only contains the relevant distinctions.
4. It provides access to dependent concepts via the independent concepts.
5. It can be easily adapted and customized.
6. It can serve as a starting point for developing ontologies for information extraction and more complex information systems.

Finally, two more comments should be made. Firstly, the limited hierarchical display does not imply that the full qualia semantics is not available for using the ontology. It simply makes a distinction between the way in which the notions are introduced and shown to a user. *Artificiality* can be specified via a **taxon,** from which it is inherited, or via explicit features that are individually introduced to more specific nodes in the hierarchy. The choice how to design the ontology and where to make the distinction is however a matter of relevance and convenience. Using the Qualia representation, there is more flexibility to design the ontological structure from a user-perspective.

Secondly, the differentiation between relevant and disjoint categories can be exploited in NLP applications. Disjoint categories should be treated in a more strict way than orthogonal stage-level categories. A query for *accessories* should be treated very differently from a query for *devices*. The term *accessories*, which carries less strict identity criteria, can refer to a much larger range of objects than *devices*. The latter terms can easily be expanded to all its hyponyms for retrieval, whereas expanding *accessories* is less obvious. In a more generic way, the customization of the hierarchies will also have an effect on basic technologies such as word-

sense-disambiguation and semantic distance measurement.

Conclusions

We described how document-based hierarchies can be extracted from documents and augmented with WordNet classifications. We also explained that the upper-levels of this classification are not directly useful. However, by imposing a more systematic and consistent Qualia classification on the WordNet top-levels, it seems possible to tune the hierarchy to the relevant distinctions. Furthermore, we can differentiate the status of the taxonomic relations on the basis of the Qualia and likewise derive principles for limiting the complexity of the hierarchies. Visualization of hierarchies can then be limited to genuinely disjoint and independent Types.

It is difficult to evaluate the appropriateness of the methodology. What makes a good hierarchy for classifying information is a subjective issue. The customization methods and programs have been developed to deliver unambiguous and minimalized hierarchies for specific customers. The time gained with the customization is very important.

References

Carslon, G. (1977) *Reference to Kinds in English.* PhD. dissertation, University of Massachusetts, Amherst.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. 423 p.

Guarino, N. 1998. *Some Ontological Principles for Designing Upper Level Lexical Resources*..Proceedings of First International Conference on Language Resources and Evaluation, Granada, 1998, pp. 527-534.

Hovy, E. (1998) *Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses*..Proceedings of First International Conference on Language Resources and Evaluation, Granada, Spain, pp. 535-542.

Justeson J.S. and S.M. Katz (1995) *Technical terminology: some linguistic properties and an algortihm for identification in text.* Natural Language Engineering, Volume 1, Part 1, March 1995, pp. 9-27.

Mihalcea R. and D.I. Moldovan (1999), *A Method for Word Sense Disambiguation of Unrestricted Text*, Proceedings of the 37th Annual Meeting of the ACL, University of Maryland, Maryland, pp. 152-158.

Pustejovsky, J. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA. 298 p.

Sanfilippo A., N. Calzolari, S. Ananiadou, R. Gaizauskas, P. Saint-Dizier, P. Vossen (eds) (1998) *EAGLES, Prelimenary Recommendations on Semantic Encoding*. Interim Report, 270 p.

Vossen, P. (ed.) (1998) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht, 1998. 251 p.

Vossen P. and L. Bloksma. (1998) *Categories and classifications in EuroWordNet*. Proceedings of First International Conference on Language Resources and Evaluation, Granada, 28-30 May 1998. pp. 399-408.

Vossen, P. (1995) *Grammatical and Conceptual Individuation in the Lexicon*, PhD. Thesis, University of Amsterdam, IFOTT, Amsterdam. 439 p.

Vossen, P. (2001fc) Extending, Trimming and Fusing WordNet for Technical Documents, Proceedings of the NAACL Workshop on Extending Wordnet, Pittsburgh, 2001.

Wilks, Y.A., B.M. Slator and L.M. Guthrie. (1996) *Electric words. Dictionaries, Computers and Meanings*. Bradford, MIT, Cambridge. p 289.