Automatically Learning Qualia Structures from the Web

Philipp Cimiano & Johanna Wenderoth
Institute AIFB
University of Karlsruhe

Abstract

Qualia Structures have many applications within computational linguistics, but currently there are no corresponding lexical resources such as WordNet or FrameNet. This paper presents an approach to automatically learn qualia structures for nominals from the World Wide Web and thus opens the possibility to explore the impact of qualia structures for natural language processing at a larger scale. Furthermore, our approach can be also used support a lexicographer in the task of manually creating a lexicon of qualia structures. The approach is based on the idea of matching certain lexico-syntactic patterns conveying a certain semantic relation on the World Wide Web using standard search engines. We evaluate our approach qualitatively by comparing our automatically learned qualia structures with the ones from the literature, but also quantitatively by presenting results of a human evaluation.

1 Introduction

Qualia Structures have been originally introduced by (Pustejovsky, 1991) and are used for a variety of purposes in Natural Language processing such as the analysis of compounds (Johnston and Busa, 1996), co-composition and coercion (Pustejovsky, 1991) as well as for bridging reference resolution (Bos et al., 1995). Further, it has also been argued that qualia structures and lexical semantic relations in general have applications in information retrieval (Voorhees, 1994; Pustejovsky et al., 1993). One major bottleneck however is that currently Qualia Structures need to be created by hand, which is probably also the reason why there are no practical system using qualia structures, but a lot of systems using globally available resources such as WordNet (Fellbaum, 1998) or FrameNet[1] as source of lexical/world knowledge. The work described in this paper addresses this issue and presents an approach to automatically learning qualia structures for nominals from the Web. The approach is inspired in recent work on using the Web to identify instances of a relation of interest such as in (Markert et al., 2003) and (Cimiano and Staab, 2004). These approaches are in essence a combination of the usage of lexico-syntactic pattens conveying a certain relation of interest such as in (Hearst, 1992), (Charniak and Berland, 1999), (Iwanska et al., 2000) or (Poesio et al., 2002) with the idea of using the web as a big corpus (Resnik and Smith, 2003), (Grefenstette, 1999), (Keller et al., 2002).

The idea of learning Qualia Structures from the Web is not only a very practical, it is in fact a principled one. While single lexicographers creating qualia structures - or lexicon entries in general - might take very subjective decisions, the structures learned from the Web do not mirror the view of a single person, but of the whole world as represented on the World Wide Web. Thus, an approach learning qualia structures from the Web is in principle more reliable than letting lexicographers craft lexical entries on their own. Obviously, on the other hand, using an automatic web based approach yields also a lot of inappropriate results which are due to 1) errors produced by the linguistic analysis (e.g. part-of-speech tagging), 2) idiosyncrasies of ranking algorithms of search machines, 3) the fact that the Web or in particular search engines are to a great extent commercially biased, 4) the fact that people also publish erroneous information on the Web, and 5) lexical ambiguities. Because of these reasons our aim is in fact not to replace lexicographers, but to support them in the task of creating qualia structures on the basis of the automatically learned qualia structures. The paper is structured as follows: Section 2 introduces qualia structures and describes the specific qualia structures we aim to acquire. Section 3 describes our approach in detail and section 4 presents a quantitative and qualitative evaluation of our approach. Before concluding, we discuss some related work in Section 5.

[1] http://framenet.icsi.berkeley.edu/

2 Qualia Structures

According to Aristotle, there are four basic factors or causes by which the nature of an object can be described (cf. (Kronlid, 2003)):

- the *material cause*, i.e. the material an object is made of
- the *agentive cause*, i.e. the source of movement, creation or change
- the *formal cause*, i.e. its form or type
- the *final cause*, i.e. its purpose, intention or aim

In his Generative Lexicon (GL) framework (Pustejovsky, 1991) reused Aristotle's basic factors for the description of the meaning of lexical elements. In fact he introduced so called *Qualia Structures* by which the meaning of a lexical element is described in terms of four roles:

- *Constitutive*: describing physical properties of an object, i.e. its weight, material as well as parts and components
- *Agentive*: describing factors involved in the *bringing about* of an object, i.e. its creator or the causal chain leading to its creation
- *Formal*: describing that properties which distinguish an object in a larger domain, i.e. orientation, magnitude, shape and dimensionality
- *Telic*: describing the purpose or function of an object

Most of the qualia structures used in (Pustejovsky, 1991) however seem to have a more restricted interpretation. In fact, in most examples the *Constitutive* role seems to describe the parts or components of an object, while the *Agentive* role is typically described by a verb denoting an action which typically brings the object in question into existence. The *Formal* role normally consists in typing information about the object, i.e. its hypernym or superconcept. Finally, the *Telic* role describes the purpose or function of an object either by a verb or nominal phrase. The qualia structure for *knife* for example could look as follows (cf. (Johnston and Busa, 1996)):

Formal:	artifact_tool
Constitutive:	blade,handle,...
Telic:	cut_act
Agentive:	make_act

Our understanding of *Qualia Structure* is in line with this restricted interpretation of the qualia roles. Our aim is to automatically acquire Qualia Structures from the Web for nominals, looking for (i) nominals describing the type of the object, (ii) verbs defining its agentive role, (iii) nominals describing its parts or components and (iv) nouns or verbs describing its intended purpose.

3 Approach

Our approach to learning qualia structures from the Web is on the one hand based on the assumption that instances of a certain semantic relation can be learned by matching certain lexico-syntactic patterns more or less reliably conveying the relation of interest in line with the seminal work of (Hearst, 1992), who defined the following patterns conveying a hypernym relation:

(1) NP_0 such as NP_1, NP_2, ..., NP_{n-1} (and|or) NP_n[2]
(2) such NP_0 as NP_1, NP_2, ... NP_{n-1} (and|or) NP_n
(3) NP_1, NP_2, ..., NP_n (and|or) other NP_0
(4) NP_0, (including|especially) NP_1, NP_2, ..., NP_{n-1} (and|or) NP_n

According to Hearst, from such patterns we can derive that for all NP_i, $1 \leq i \leq n$, $hypernym(NP_i, NP_0)$. For example, for the expression: *Bruises, wounds, broken bones or other injuries*, we would extract: *hypernym(bruise,injury), hypernym(broken bone,injury)* and *hypernym(wound,injury)*. However, it is well known that Hearst-style patterns occur rarely, such that it seems intuitive to match them on the Web. So in our case we are looking not only for the hypernym relation (comparable to the *Formal*-Relation) but for similar patterns conveying a *Constitutive*, *Telic* or *Agentive* relation. As currently there is no support for searching using regular expressions in standard search engines such as Google or Altavista[3], our approach consists of 5 phases (compare Figure 1):

1. generate for each qualia role a set of so called *clues*, i.e. search engine queries indicating the relation of interest
2. download the snippets of the 10 first Google hits matching the generated clues [4]
3. part-of-speech-tagging of the downloaded snippets
4. match regular expressions conveying the qualia role of interest
5. weight the returned qualia elements according to some measure

The outcome of this process are then so called *Weighted Qualia Structures* (WQSs) in which every

[2] NP_i stands for a noun phrase.

[3] An exception is certainly the Linguist's Search Engine (Resnik and Elkiss, 2003)

[4] The reason for using only the 10 first hits is to maintain efficiency. With the current setting the systems needs between 3 and 10 minutes to generate the qualia structure for a given nominal

qualia element in a certain role is weighted according to some measure. The patterns in our pattern library are actually tuples (p, c) where p is a regular expression defined over part-of-speech tags and c a function $c : string \rightarrow string$ called the *clue*. Given a nominal t and a clue c, the query $c(t)$ is sent to the Google API and we download the abstracts of the first n documents matching this query and then process the abstracts to find instances of pattern p. For example, given the clue $f(x) = "such\ as\ "\pi(x)$ and the instance *computer* we would download n abstracts matching the query f(computer), i.e. "such as computers". Hereby $\pi(x)$ is a function returning the plural form of x. We implemented this function as a lookup in a lexicon in which plural nouns are mapped to their base form. With the use of such clues, we thus download a number of Google-abstracts in which a corresponding pattern will probably be matched thus restricting the linguistic analysis to a few promising pages. The downloaded abstracts are then part-of-speech tagged using QTag (Tufis and Mason, 1998). Then we match the corresponding pattern p in the downloaded snippets thus yielding candidate qualia elements as output. In our approach we then calculate the weight of a candidate qualia element e for the term t we want to compute the qualia structure for by the *Jaccard Coefficient*:

$$\frac{GoogleHits(e+t)}{GoogleHits(e) + GoogleHits(t) - GoogleHits(e+t)}$$

The result is then a *Weighted Qualia Structure* (WQS) in which for each role the qualia elements are weighted according to this Jaccard coefficient. In what follows we describe in detail the procedure for acquiring qualia elements for each qualia role. In particular, we describe in detail the clues and lexico-syntactic patterns used. In general, the patterns have been crafted by hand, testing and refining them in an iterative process, paying attention to maximize their coverage but also accuracy.
In general it is important to mention that by this approach we are not able to detect and separate multiple meanings of words, i.e. to handle polysemy, which is appropriately accounted for in the framework of the Generative Lexicon (Pustejovsky, 1991).

3.1 The Formal Role

To derive qualia elements for the *Formal* role, we first download for each of the clues in Table 1 the first 10 abstracts matching the clue and then process them offline matching the patterns defined over part-of-speech-tags[5] thus yielding up to 10 different qualia element candidates per clue. The patterns are specified in form of regular expressions, whereby the part-of-speech tags are always

[5]We use the well-known Penn Treebank tagset described at http://www.computing.dcu.ie/∼acahill/tagset.html.

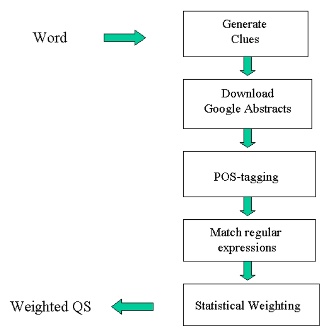

Figure 1: General Approach

given in square brackets after the token. Further, besides using the traditional regular expression operators such as $+$, $*$ and ?, we also use Perl-like symbols such as $\backslash w$ denoting any alphabetic character as well as [a-z] denoting the set of all lower case letters.
As there are 4 different clues for the *Formal* role, we thus yield up to 40 qualia elements as potential candidates to fill the *Formal* role. In general, we paid attention to create clues relying on indefinite articles as we found out that they produce more general and reliable results than when using definite articles. In order to choose the correct indefinite article – *a* or *an* – or even using no article at all, we implemented some ad-hoc heuristics checking if the first letter of the term in question is a vowel and checking if the term is used more often with an article or without an article on the Web by a set of corresponding Google queries. The alternative '(a/an/?)' means that we use either the indefinite article 'a' 'an' or no article depending on the results of the above mentioned Google queries.
A general question raised also by Hearst (Hearst, 1992) is how to deal with NP modification. Hearst's conclusion is that this depends on the application. In our case we mainly remove adjective modifiers, keeping only the heads of noun phrases as candidate qualia elements. The lemmatized heads of the NP_F noun phrase are then regarded as qualia role candidates for the *Formal* role. These candidates are then weighted using the above defined *Jaccard Coefficient* measure. Hereby, a noun phrase is an instance matching the following regular expression:

NP:=[a-z]+[DT]? ([a-z]+[JJ])+? <u>([a-z]+[NN(S?)])+</u>,

where the head is the underlined expression, which is lemmatized and considered as a candidate qualia element. After some initial experiments we decided not to use the patterns 'X is Y' and 'X is a kind of Y' such as in *a book is an item* or *a book is a kind of publication* as well as the pattern 'Y, including X' (compare (Hearst, 1992)) as we found that in our settings they delivered quite spurious results.

Clue	Pattern
such as $\pi(t)$	NP$_F$,? such[DT] as[IN] NP
especially $\pi(t)$	NP$_F$,? especially[RB] NP
$\pi(t)$ or other	NP or[CC] other[JJ] NP$_F$
$\pi(t)$ and other	NP and[CC] other[JJ] NP$_F$

Table 1: Clues and Patterns for the *Formal* role

3.2 The Constitutive Role

The procedure for finding elements of the *Constitutive* role is similar to the one described above for the *Formal* role. The corresponding clues and patterns are given in Table 2. As above, the candidate qualia elements are then the lemmatized heads of the noun phrase NP$_C$.

Clue	Pattern
(a/an)? t is made up of	NP is[VBZ] made[VBN] up[RP] of[IN] NP$_C$
$\pi(t)$ are made up of	NP are[VBP] made[VBN] up[RP] of[IN] NP$_C$
(a/an)? t is made of	NP are[VBP] made[VBN] of[IN] NP$_C$
$\pi(t)$ are made of	NP are[VBP] made[VBN] of[IN] NP$_C$
(a/an)? t comprises	NP comprises[VBZ] NP$_C$
$\pi(t)$ comprise	NP comprise[VBP] NP$_C$
(a/an)? t consists of	NP consists[VBZ] of[IN] NP$_C$
$\pi(t)$ consist of	NP consist[VBP] of[IN] NP$_C$

Table 2: Clues and Patterns for the *Constitutive* Role

As an additional heuristic, we test if the lemmatized head of NP$_C$ is an element of the following list containing nouns denoting an indication of amount: {*variety, bundle, majority, thousands, million, millions, hundreds, number, numbers, set, sets, series, range*} and furthermore this NP$_C$ is followed by the preposition 'of'. In that case we would take the head of the noun phrase after the preposition 'of' as potential candidate of the *Constitutive* role. For example, when considering *a conversation is made up of a series of observable interpersonal exchanges*, we would take *exchange* as a potential qualia element candidate instead of *series*.

3.3 The Telic Role

The *Telic* Role is in principle acquired in the same way as the *Formal* and *Constitutive* roles with the exception that the qualia element is not only the head of a noun phrase, but also a verb or a verb followed by a noun phrase. Table 3 gives the corresponding clues and patterns. In particular, the returned candidate qualia elements are the lemmatized underlined expressions in PURP:=\w+[VB] NP | NP | be[VB] \w+[VBD]).

Clue	Pattern
purpose of a t is	purpose[NN] of[IN] NP$_0$ is[VBZ] (to[TO])? PURP
purpose of $\pi(t)$ is	purpose[NN] of[IN] NP$_0$ is[VBZ] (to[TO])? PURP
(a/an)? t is used to	(A|a|An|an) NP$_0$ is[VBZ] used[VBN] to[TO] PURP
$\pi(t)$ are used to	NP$_0$ are[VBZ] used[VBN] to[TO] PURP

Table 3: Clues and Patterns for the *Telic* Role

3.4 The Agentive Role

As mentioned in (Hearst, 1992), it is not always as straightforward to find lexico-syntactic patterns reliably conveying a certain relation. In fact, we did not find any patterns reliably identifying qualia elements for the *Agentive* role. Certainly, it would have been possible to find the source of the creation by using patterns such as *X is made by Y* or *X is produced by Y*. However, we found that these patterns do not reliably convey a verb describing how an object is brought into existence. The fact that it is far from straightforward to find patterns indicating an *Agentive* role is further corroborated by the research in (Yamada and Baldwin, 2004), in which only one pattern indicating a qualia relation is used, namely 'NN BE V[+en]' in order to match passive constructions such as *the book was written*. On the other hand it is clear that constructing a reliable clue for this pattern is not straightforward given the current state-of-the-art concerning search engine queries. Nevertheless, in order to also get results for the *Agentive* role, we apply a different method here. Instead of issuing a query which is used to search for possible candidates for the role, we take advantage of the fact that the verbs which describe how something comes into being, particularly artificial things, are often quite general phrases like "make, produce, write, build...". So instead of generating clues as above, we calculate the value $\frac{GoogleHits(<AGENTIVE\_VERB>\ a\ t)}{GoogleHits(t)}$ for the nominal we want to acquire a qualia structure for as well as the following verbs: *build, produce, make, write, plant, elect, create, cook, construct* and *design*. If this value is over a threshold (0.0005 in our case), we assume that it is a valid filler of the *Agentive* qualia role.

4 Evaluation

We evaluate our approach for the lexical elements *knife, beer, book*, which are also discussed in (Johnston and

Busa, 1996) or (Pustejovsky, 1991), as well as *computer*, an abstract noun, i.e. *conversation*, as well as two very specific multi-term words, i.e. *natural language processing* and *data mining*. We give the automatically learned weighted Qualia Structures for these entries in Figures 3, 4, 5 and 6. The evaluation of our approach consists on the one hand of a discussion of the weighted qualia structures, in particular comparing them to the ideal structures form the literature. On the other hand, we also asked a student at our institute to assign credits to each of the qualia elements from 0 (incorrect) to 3 (totally correct) whereby 1 credit meaning 'not totally wrong' and 2 meaning 'still acceptable'.

4.1 Quantitative Evaluation

The distribution of credits for each qualia role and term is given in Table 4. It can be seen that with three exceptions: *beer→formal*, *book→agentive* as well as *beer→constitutive*, '3' is the mark assigned in most cases to the automatically learned qualia elements. Further, for almost every query term and qualia role, at least 50% of the automatically learned qualia structures have a mark of '2' or '3' – the only exceptions being *beer→formal* with 45.45%, *book→agentive* with 33.33% and *beer→constitutive* with 28.57%. In general this shows that the automatically learned qualia roles are indeed reasonable. Considering the average over all the terms ('All' in the table), we observe that the qualia role which is recognized most reliably is the *Telic* one with 73.15% assignments of credit '3' and 75.93% of credits '2' or '3', followed by the *Agentive* role with 71.43% assignments of credit 3. The results for the *Formal* and *Constitutive* role are still reasonable with 62.09% assignments of credit '3' and 66.01% assignments of credits '2' or '3' for the *Formal role*; and respectively 61.61% and 64.61% for the *Constitutive* role. The worst results are achieved for the *Constitutive* role due to the fact that 26.26% of the qualia elements are regarded as totally wrong. Table 5 supports the above claims and shows the average credits assigned by the human evaluator per query term and role. It shows again that the roles with the best results are the *Agentive* and *Telic* roles, while the *Formal* and *Constitutive* roles are not identified as accurately. This is certainly due to the fact that the patterns for the *Telic* role are much less ambiguous than the ones for the *Formal* and *Constitutive* roles. Finally, we also discuss the correlation between the credits assigned and the *Jaccard Coefficient*. Figure 2 shows this correlation. While for the *Formal* role the correlation is as expected, i.e. the higher the credit assigned, the higher also the Jaccard Coefficient, for the *Constitutive* and *Telic* roles this correlation is unfortunately less clear, thus making the task of finding a cut-off threshold more difficult.

4.2 Qualitative Evaluation & Discussion

In this section we provide a more subjective evaluation of the automatically learned qualia structures by comparing them to ideal qualia structures discussed in the literature wherever possible. In particular, we discuss more in detail the qualia structure for *book*, *knife* and *beer* and leave the detailed assessment of the qualia structures for *computer*, *natural language processing*, *data mining* and *conversation* to the interested reader.

For book, the first four candidates of the *Formal* role, i.e. *product, item, publication* and *document* are very appropriate, but alluding to the *physical object* meaning of book as opposed to the meaning in the sense of *information container* (compare (Pustejovsky, 1991). As candidates for the *Agentive* role we have *make, write* and *create* which are appropriate, *write* being the ideal filler of the *Agentive* role according to (Pustejovsky, 1991). For the *Constitutive* role of *book* we get – besides *it* at the first position which could be easily filtered out – *sign* (2nd position), *letter* (3rd position) and *page* (6th position), which are quite appropriate. The top four candidates for the *Telic* role are *give, select, read* and *purchase*. It seems that *give* is emphasizing the role of a book as a gift, *read* is referring to the most obvious purpose of a book as specified in the ideal qualia structures of (Pustejovsky, 1991) as well as (Johnston and Busa, 1996) and *purchase* denotes the more general purpose of a book, i.e. to be bought.

The first element of the *Formal* role of *knife* unfortunately denotes the material it is typically made of, i.e. *steel*, but the next 5 elements are definitely appropriate: *weapon*, *item*, *kitchenware*, *object* and *instrument*. The ideal element *artifact_tool* (compare (Johnston and Busa, 1996)) can be found at the 10th position. The results are interesting in that on the one hand the most prominent meaning of *knife* according to the web is the one of a *weapon*. On the other hand our results are more specific, classifying a knife as *kitchenware* instead of merely as an *artifact_tool*. Very interesting are the specific and accurate results at the end of the list. The reason why they appear at the end is that the Jaccard Coefficient ranks them lower because they are more specific, thus appearing less frequently. This shows that using some other measure less sensitive to frequency could yield more accurate results. The fillers of the *Agentive* role *produce, make* and *create* seem all appropriate, whereby *make* corresponds exactly to the ideal filler for the *Agentive* role as mentioned in (Johnston and Busa, 1996). The results for the *Constitutive* role contain not only parts but also materials a knife is made of and thus contain more information than the typical qualia structures assumed in the literature. The best results are (in this order) *blade, metal, steel, wood* and *handle* at the 6th position. In fact, in the ideal qualia structure in (Johnston and Busa, 1996) *blade* and *han-*

	Formal			
	0	1	2	3
Book	2/17 (11.76%)	4/17 (23.52%)	1/17 (5.88%)	10/17 (58.82%)
Computer	8/28 (28.57%)	1/28 (3.57%)	2/28 (7.14%)	17/28 (60.71%)
Knife	3/16 (18.75%)	0/16 (0%)	0/16 (0%)	13/16 (81.25%)
Beer	12/22 (54.54%)	0/22 (0%)	2/22 (9.09%)	8/22 (36.36%)
Data Mining	6/25 (24%)	0/25 (0%)	0/25 (0%)	19/25 (76%)
Natural Language Processing	2/15 (13.33%)	1/15 (6.66%)	0/15 (0%)	12/15 (80%)
Conversation	10/30 (33.33%)	4/30 (13.33%)	0/30 (0%)	16/30 (53.33%)
All	43/153 (28.10%)	11/153 (7.19%)	6/153 (3.92%)	95/153 (62.09%)
	Agentive			
Book	0/3 (0%)	2/3 (66.66%)	0/3 (0%)	1/3 (33.33%)
Computer	0/1 (0%)	0/1 (0%)	0/1 (0%)	1/1 (100%)
Knife	0/3 (0%)	0/3 (0%)	0/3 (0%)	3/3 (100%)
Beer	0/3 (0%)	1/3 (33.33%)	0/3 (0%)	2/3 (66.66%)
Data Mining	0/1 (0%)	0/1 (0%)	0/1 (0%)	1/1 (100%)
Natural Language Processing	0/1 (0%)	0/1 (0%)	0/1 (0%)	1/1 (100%)
Conversation	1/2 (50%)	0/2 (0%)	0/2 (0%)	1/2 (50%)
All	1/14 (7.14%)	3/14 (21.43%)	0/14 (0%)	10/14 (71.43%)
	Constitutive			
Book	8/29 (27.58%)	4/29 (13.79%)	1/29 (3.44%)	16/29 (55.17%)
Computer	6/26 (23.07%)	1/26 (3.84%)	0/26 (0%)	19/26 (73.07%)
Knife	4/15 (26.66%)	0/15 (0%)	0/15 (0%)	11/15 (73.33%)
Beer	5/7 (71.42%)	0/7 (0%)	0/7 (0%)	2/7 (28.57%)
Data Mining	0/1 (0%)	0/1 (0%)	0/1 (0%)	1/1 (100%)
Natural Language Processing				
Conversation	3/21 (14.28%)	4/21 (19.04%)	0/21 (0%)	14/21 (66.66%)
All	26/99 (26.26%)	9/99 (9%)	3/99 (3%)	61/99 (61.61%)
	Telic			
Book	3/22 (13.63%)	2/22 (9.09%)	3/22 (13.63%)	14/22 (63.63%)
Computer	0/27 (0%)	3/27 (11.11%)	0/27 (0%)	24/27 (88.88%)
Knife	5/18 (27.77%)	0/18 (0%)	0/18 (0%)	13/18 (72.22%)
Beer				
Data Mining	2/22 (9.09%)	4/22 (18.18%)	0/22 (0%)	16/22 (72.72%)
Natural Language Processing	1/6 (16.66%)	0/6 (0%)	0/6 (0%)	5/6 (83.33%)
Conversation	6/13 (46.15%)	0/13 (0%)	0/13 (0%)	7/13 (53.84%)
All	17/108 (15.74%)	9/108 (8.33%)	3/108 (2.78%)	79/108 (73.15%)

Table 4: Distribution of credits for each role and term

	Formal	Agentive	Constitutive	Telic
Book	2.12	1.67	1.86	2.27
Computer	2	3	2.23	2.78
Knife	2.44	3	2.2	2.17
Beer	1.27	2.33	0.96	n.a.
Data Mining	2.28	3	3	2.36
Natural Language Processing	2.47	3	n.a.	2.5
Conversation	1.73	1.5	2.19	1.62
All	1.99	2.36	2.02	2.33

Table 5: Average credits for each role

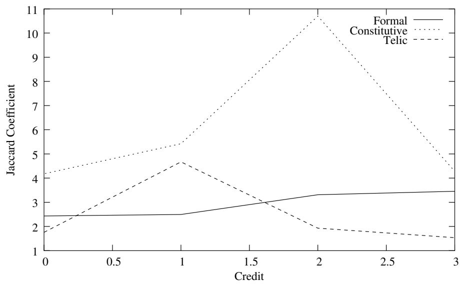

Figure 2: Average Jaccard Coefficient value per credit

dle are mentioned as fillers of the *Constitutive* role, while there are no elements describing the materials of which a knife is made of. Finally, the top four candidates for the *Telic* role are *kill, slit, cut* and *slice*, whereby *cut* corresponds to the ideal filler of the qualia structure for *knife* as mentioned in (Johnston and Busa, 1996).

Considering the qualia structure for *beer*, it is surprising that no purpose has been found. The reason is that currently no results are returned by Google for the clue *a beer is used to* and the four snippets returned for *the purpose of a beer* contain expressions of the form *the purpose of a beer is to drink it* which is not matched by our patterns as *it* is a pronoun and not matched by our NP pattern (unless it is matched by an error as in the Qualia Structure for *book* in Figure 4). Considering the results for the *Formal* role, the elements *drink* (1st), *alcohol* (2nd) and *beverage* (4th) are much more specific than *liquid* as given in (Pustejovsky, 1991), while *thing* at the 3rd position is certainly too general. Furthermore, according to the automatically learned qualia structure, *beer* is made of *rice*, *malt* and *hop*, which are perfectly reasonable results. Very interesting are the results *concoction* and *libation* for the *Formal* role of beer, which unfortunately were rated low by our evaluator (compare Figure 3).

Overall, the discussion has shown that the results produced by our method are reasonable when compared to the qualia structures from the literature. In general, our method produces in some cases additional qualia candidates, such as the ones describing the material a knife is typically made of. In other cases it discovers more specific candidates, such as for example *weapon* or *kitchenware* as elements of the *Formal* role for knife instead of the general term *artifact_tool*.

5 Related Work

There is quite a lot of work related to the use of linguistic patterns to discover certain ontological relations from text. Hearst's (Hearst, 1992) seminal work had the aim of discovering taxonomic relations from electronic dictionaries. The precision of the *is-a*-relations learned is 61/106 (57.55%) when measured against WordNet as gold standard, which is comparable to our results. Hearst's idea has been reapplied by different researchers with either slight variations in the patterns used (Iwanska et al., 2000), to acquire knowledge for anaphora resolution (Poesio et al., 2002), or to discover other kinds of semantic relations such as part-of relations (Charniak and Berland, 1999) or causation relations (Girju and Moldovan, 2002).

Instead of matching these patterns in a large text collection, some researchers have recently turned to the Web to match these patterns such as in (Cimiano and Staab, 2004) or (Markert et al., 2003). (Cimiano and Staab, 2004) for example aim at learning instance-of as well as taxonomic (*is-a*) relations. This is very related to the acquisition of the *Formal* role proposed here. (Markert et al., 2003) aim at acquiring knowledge for anaphora resolution, while (Etzioni et al., 2004) aim at learning the complete extension of a certain concept. For example, they aim at finding all the actors in the world.

Our approach goes further in that it not only learns typing, superconcept or instance-of relations, but also *Constitutive* and *Telic* relations.

There also exist approaches specifically aiming at learning qualia elements from corpora based on machine learning techniques. (Claveau et al., 2003) for example use Inductive Logic Programming to learn if a given verb is a qualia element or not. However, their approach goes not as far as learning the complete qualia structure for a lexical element in an unsupervised way as presented in our approach. In fact, in their approach they do not distinguish between different qualia roles and restrict themselves to verbs as potential fillers of qualia roles. (Yamada and Baldwin, 2004) present an approach to learning *Telic* and *Agentive* relations from corpora analyzing two different approaches: one relying on matching certain lexico-syntactic patterns as in the work presented here, but also a second approach consisting in training a maximum entropy model classifier. Their conclusion is that the results produced by the classification approach correlate better with two hand-crafted gold standards.

The patterns used by (Yamada and Baldwin, 2004) differ substantially from the ones used in this paper, which is mainly due to the fact that search engines do not provide support for regular expressions and thus instantiating a pattern as 'V[+ing] Noun' is impossible in our approach as the verbs are unknown a priori.
Finally, (Pustejovsky et al., 1993) present an interesting framework for the acquisition of semantic relations from corpora not only relying on statistics, but guided by theoretical lexicon principles.

6 Conclusion

We have presented an approach to automatically learning Qualia Structures from the Web. Such an approach is especially interesting either for lexicographers aiming at constructing lexicons, but even more for natural language processing systems relying on deep lexical knowledge as represented by qualia structures. We have in particular shown that the qualia structures learned by our system are reasonable. In general, it is valid to claim that our system is the first one automatically producing complete qualia structures for a given nominal.
Our system can be tested online at http://km.aifb.uni-karlsruhe.de/pankow/qualia/. Further work will aim at improving the system but also at using the automatically learned structures within NLP applications.

Acknowledgments The work reported in this paper has been partially supported by the SmartWeb project[6], funded by the German Ministry of Research. Thanks also to Laura Goebes for assisting in the evaluation of the system.

[6]http://www.smartweb-projekt.de/

Knife		
Formal		
steel	3.8666	3
weapon	3.4876	3
item	1.7458	3
kitchenware	1.6840	3
object	1.6025	3
instrument	1.2963	3
utensil	1.2886	3
court	1.1441	0
equipment	0.9479	3
tool	0.7090	3
action	0.7028	0
time	0.6590	0
cutting instrument	0.0739	3
cutting instruments	0.0551	3
emergency items	0.0383	3
cutting weapons	0.0232	3
Agentive		
produce		3
make		3
create		3
Constitutive		
blade	5.4618	3
metal	5.0205	3
steel	3.8666	3
wood	2.9699	3
person	2.6829	0
handle	1.9223	3
tang	1.6784	3
gold	1.6609	0
alloy	1.2466	3
dragonfly	0.8742	3
model	0.7513	3
tool	0.7090	0
quality	0.6575	3
group	0.5764	0
rotating discs	0.0062	3
Telic		
kill	3.7626	3
slit	3.4829	3
cut	3.4373	3
slice	3.2499	3
begin	2.4192	0
split	1.7241	3
avoid	1.3190	0
score	1.0204	0
an instrument	0.8137	0
process	0.5327	3
prune	0.4505	3
incise	0.0573	3
cut things	0.0545	3
remove moisture	0.0479	3
add details	0.0361	0
cut a flap	0.0264	3
split a cake	0.0010	3
slit a wide variety	0.0004	3

Beer		
Formal		
drink	9.6677	3
alcohol	4.6006	3
thing	4.0028	3
beverage	3.6182	3
adventure	3.0825	0
mistake	2.7014	0
matter	2.6533	0
style	2.1583	0
delight	1.9198	3
people	1.4465	0
creation	1.2201	0
can	0.9433	3
list	0.8432	0
product	0.8224	3
refreshment	0.5328	3
concoction	0.4851	0
libation	0.1147	0
summery	0.0872	0
adult beverages	0.0848	2
speciality beers	0.0269	2
looney things	0.0002	0
Agentive		
produce		3
make		3
create		1
Constitutive		
rice	2.9871	0
malt	2.5724	3
hop	2.1744	3
bottom	2.1179	0
continuum	0.4808	0
puree	0.3563	0
stoneware	0.3325	0

Figure 3: Weighted Qualia Structure for *knife* and *beer*

References

J. Bos, P. Buitelaar, and M. Mineur. 1995. Bridging as coercive accomodation. In E. Klein, S. Manandhar, W. Nutt, and J. Siekmann, editors, *Working Notes of the Edinburgh Conference on Computational Logic and Natural Language Processing (CLNLP-95)*.

E. Charniak and M. Berland. 1999. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 57–64.

P. Cimiano and S. Staab. 2004. Learning by googling. *SIGKDD Explorations*, 6(2), December.

V. Claveau, P. Sebillot, C. Fabre, and P. Bouillon. 2003. Learning semantic lexicons from a part-of-speech and semantically tagged corpus using inductive logic programming. *Journal of Machine Learning Research*, (4):493–525.

O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. 2004. Web-scale information extraction in KnowItAll (preliminary results). In *Proceedings of the 13th World Wide Web Conference*, pages 100–109.

C. Fellbaum. 1998. *WordNet, an electronic lexical database*. MIT Press.

R. Girju and M. Moldovan. 2002. Text mining for causal relations. In *Proceedings of the FLAIRS Conference*, pages 360–364.

G. Grefenstette. 1999. The WWW as a resource for example-based MT tasks. In *Proceedings of ASLIB'99 Translating and the Computer 21*.

M.A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th In-*

Book		
Formal		
product	34.6238	3
item	33.8573	3
publication	20.2621	3
document	14.4778	3
history	12.7262	1
project	8.9809	2
material	8.6704	3
reader	8.3890	0
resource	7.7259	3
source	7.6739	3
piece	7.6131	3
format	7.2203	0
tool	6.1124	1
object	3.7705	3
specifics	0.5374	1
library materials	0.1468	3
library property	0.0026	1
Agentive		
make		1
write		3
create		1
Constitutive		
it	21.5785	0
sign	21.0870	3
letter	18.7778	3
part	11.7830	1
individual	11.4043	0
page	10.9202	3
collection	10.7901	0
teaching	10.7004	2
language	9.6041	1
period	9.4002	0
paper	9.3551	3
table	8.7089	3
material	8.6704	3
word	8.1424	3
piece	7.6131	0
chapter	7.4746	3
presentation	7.0955	3
detail	6.8218	3
minute	5.3550	0
sheet	4.4369	3
lie	3.0866	1
ticket	2.3198	0
ink	2.2769	3
dot	1.7427	3
leather	1.1162	1
leaf	1.0266	3
title page	0.3639	3
peice	0.0530	0
dedication page	0.0076	3
Telic		
give	14.8954	1
select	12.9594	0
read	12.4937	3
purchase	9.0372	3
support	8.0204	3
identify	7.9388	1
represent	5.7829	2
inspire	1.7292	3
convey	1.3940	3
present information	0.0728	3
provide additional information	0.0368	3
convey information	0.0260	3
filch	0.0101	3
share a story	0.0081	3
commit crime	0.0061	0
contain words	0.0055	3
introduce concepts	0.0038	2
traprock	0.0015	0
stock libraries	0.0009	3
hold a collection	0.0008	3
fund special projects	0.0007	2
support teachings	0.0001	3

Computer		
Formal		
technology	20.3667	3
information	20.2418	0
network	14.8052	3
hardware	14.6539	3
service	13.9161	3
office	12.2881	0
equipment	7.4594	2
machine	7.0099	3
item	6.7469	3
device	5.6259	3
medium	4.0503	3
fix	3.9188	0
piece	3.5898	3
notebook	2.1126	3
circuit	1.8663	0
consumer electronics	1.1544	0
appliance	1.0045	3
toy	0.7934	3
office equipment	0.4055	3
datum	0.3262	0
computer clipart	0.3156	1
mentality	0.1158	0
network device	0.0343	3
artefact	0.0339	3
data stores	0.0133	3
display screen equipment	0.0042	2
library equipment	0.0037	3
complex computer processes	0.0001	0
Agentive		
build		3
Constitutive		
software	25.5230	3
hardware	14.6539	3
part	14.6224	1
electronics	9.6139	3
individual	9.3791	0
memory	8.9683	3
man	5.9584	0
device	5.6259	3
unit	5.2078	3
component	4.3808	3
switch	4.2159	3
mix	3.8996	0
string	1.8896	3
circuit	1.8663	0
silicon	1.7717	3
actor	1.2127	0
processing unit	0.1444	3
individual components	0.1122	3
hardware components	0.1087	3
centra	0.0530	0
computer codes	0.0463	3
plastic case	0.0167	3
data storage device	0.0077	3
transitors	0.0022	3
Telic		
make	16.9616	1
access	15.5691	3
control	12.2216	3
run	8.6411	3
assist	4.1410	3
publish	3.0015	3
solve	2.9701	3
facilitate	2.8860	3
insight	2.2718	3
combine	1.9592	1
calculate	1.2977	3
execute	1.2792	3
translate	1.2530	3
suppose	1.1340	3
provide information	0.8969	3
access data	0.1025	3
imitate	0.0998	1
provide feedback	0.0900	3
human freedom	0.0065	3
teach children	0.0266	3
enable people	0.0255	3
manage information	0.0231	3
process words	0.0009	3
support program goals	0.0003	3
reduce analysis time	0.0002	3
perform useful computations	0.0001	3

Conversation		
Formal		
concept	6.6834	3
expression	5.8487	3
context	5.2338	3
object	4.6343	0
sound	4.4566	0
function	4.1414	0
material	4.1324	0
place	3.7806	0
employee	3.4710	0
skill	3.3323	3
interaction	3.1092	3
communication	3.0006	3
activity	2.9859	3
people	2.9027	0
label	2.7427	3
time	2.6158	1
source	1.6782	0
text	1.5877	1
transmission	1.2251	3
information	1.2182	3
contact	1.1309	3
utterance	0.9499	1
transaction	0.9412	3
school activities	0.2094	3
datum	0.1462	3
mannerism	0.0635	0
communication difficulties	0.0412	1
ambient audio	0.0148	3
official forms	0.0140	3
priceless tidbits	0.0002	0
Agentive		
make		3
create		0
Constitutive		
relationship	6.1848	3
silence	5.7213	3
answer	5.6855	3
question	4.8714	3
sentence	4.8663	3
story	4.4669	3
laughter	3.1766	1
unit	2.9359	1
tree	2.7633	0
contribution	2.6421	3
world	2.1804	0
sequence	1.8986	3
requests	1.4969	3
repetition	1.4267	3
token	1.2746	1
bonus	1.2155	1
pauses	1.1568	3
utterance	0.9499	0
cliches	0.2556	3
interpersonal exchanges	0.0082	3
brief debates	0.0003	3
Telic		
exchange	4.2769	3
establish	3.3530	3
further	3.2694	0
allow	3.2489	3
create	2.7141	0
generate	2.0107	0
get	1.9484	0
gloss	0.4780	0
exchange information	0.2313	3
exchange ideas	0.1896	3
enable people	0.1151	3
pass time	0.0469	0
teach skills	0.0171	3

Figure 4: Weighted Qualia Structures for *book, computer* and *conversation*

Data Mining		
Formal		
data analysis	2.1492	3
intelligence	1.4242	0
analysis	1.2009	3
tool	1.1987	3
prediction	0.9682	3
approach	0.7279	3
speciality	0.6245	3
system	0.6018	3
application	0.5209	3
functionality	0.3974	3
process	0.3840	3
mechanism	0.3503	3
type	0.3372	0
practice	0.3310	3
technology	0.3240	3
activity	0.3207	3
employment	0.2565	0
use	0.2128	3
name	0.1944	3
area	0.1856	0
datum	0.1701	0
data warehousing technologies	0.1497	3
subject	0.1403	0
information process	0.0498	3
information process techniques	0.0005	3
Agentive		
design		3
Constitutive		
knowledge	0.7062	3
Telic		
connect	0.5949	0
achieve	0.3651	3
uncover	0.3460	3
research	0.3374	3
answer	0.2122	3
support	0.2025	3
look	0.1834	0
provide information	0.1527	3
search	0.1451	3
tell	0.1099	1
identify patterns	0.0959	3
discover patterns	0.0934	3
identify trends	0.0765	3
provide a foundation	0.0620	1
improve services	0.0559	3
gain business intelligence	0.0048	3
explore knowledge	0.0045	3
detect dependencies	0.0036	3
gain business	0.0223	1
analyse large volumes	0.0022	1
find new prospects	0.0011	3
analyze disparate customer data	0.0002	3

Figure 5: Weighted Qualia Structure for *data mining*

Natural Language Processing		
Formal		
linguistics	1.0047	3
technique	0.4983	3
intelligence	0.3559	3
method	0.2748	3
model	0.1847	3
aspect	0.1380	3
scheme	0.1258	3
system	0.0750	1
research	0.0636	3
application	0.0603	3
science	0.0536	3
technology	0.0414	3
area	0.0373	0
product	0.0337	0
document processing applications	0.0174	3
Agentive		
design		3
Constitutive		
Telic		
build	0.1037	3
keep track	0.0820	3
understand	0.0662	3
soften	0.0501	0
provide	0.0384	3
build tailored knowledge base	0.0008	3

Figure 6: Weighted Qualia Structure for *natural language processing*

ternational Conference on Computational Linguistics, pages 539–545.

L.M. Iwanska, N. Mata, and K. Kruger. 2000. Fully automatic acquisition of taxonomic knowledge from large corpora of texts. In L.M. Iwanksa and S.C. Shapiro, editors, *Natural Language Processing and Knowledge Processing*, pages 335–345. MIT/AAAI Press.

M. Johnston and F. Busa. 1996. Qualia structure and the compositional interpretation of compounds.

F. Keller, M. Lapata, and O. Ourioupina. 2002. Using the web to overcome data sparseness. In *Proceedings of EMNLP-02*, pages 230–237.

F. Kronlid. 2003. Modes of explanation - aristotelian philosophy and pustejovskyan linguistics. Ms. University of Gteborg.

K. Markert, N. Modjeska, and M. Nissim. 2003. Using the web for nominal anaphora resolution. In *EACL Workshop on the Computational Treatment of Anaphora*.

M. Poesio, T. Ishikawa, S. Schulte im Walde, and R. Viera. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*.

J. Pustejovsky, P. Anick, and S. Bergler. 1993. Lexical semantic techniques for corpus analysis. *Computational Lingustics, Special Issue on Using Large Corpora II*, 19(2):331–358.

J. Pustejovsky. 1991. The generative lexicon. *Computational Linguistics*, 17(4):209–441.

P. Resnik and A. Elkiss. 2003. The linguist's search engine: Getting started guide. Technical Report LAMP-TR-108/CS-TR-4541/UMIACS-TR-2003-109, University of Maryland, College Park, November.

P. Resnik and N. Smith. 2003. The web as a parallel corpus. *Computational Lingusitics*, 29(3):349–380.

D. Tufis and O. Mason. 1998. Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, pages 589–96.

E.M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69.

I. Yamada and T. Baldwin. 2004. Automatic discovery of telic and agentive roles from corpus data. In *Proceedings of the The 18th Pacific Asia Conference on Language, Information and Computation (PACLIC 18)*.