Daniel Berndt

Qualitative Enhancements by Quantitative Analysis
— Explaining Corpus based Phenomena —

First Supervisor: Louise McNally
Second Supervisor: Stefan Evert
Date of Submission: August 27, 2009
Abstract

This thesis emphasises the importance of a solid quantitative analysis for investigations on corpora in order to gain a profound insight into features that constitute the phenomena at hand. To exemplify this, I present my research on key elements that trigger the way of how nation-noun expressions are constructed and I will reveal qualitative explanations for quantitative findings.
Contents

1 Introduction 4

2 Motivation and Methods 4
  2.1 The primal Problem 5
  2.2 The underlying Data 5
  2.3 Analysing that Data 7
  2.4 A deeper Insight into the Methods 10
    2.4.1 Queries on the British National Corpus 10
    2.4.2 WordNet and Top Concept Ontology 11
    2.4.3 Entropy as a Means of measuring Specificity 13

3 The Strange Plot 15
  3.1 Origin of the Strange Plot 16
  3.2 The Strange Plot from another Point of View 17
  3.3 Finding Hypotheses explaining the Strange Plot 18

4 Advance towards the Core of the Strange Plot 18
  4.1 Type based Analysis 18
  4.2 Gained Insights 19
  4.3 Analysing the new Approach 21

5 Advance towards the Core of the primal Problem 23
  5.1 Entropy Distribution 23
  5.2 Frequency Distribution 25
  5.3 Correlation of Entropy and Frequency 25
  5.4 Further Reflections concerning Entropy 28
  5.5 Final Considerations 30

6 Conclusion and Outlook 32

7 Acknowledgements 33

A Feature Hierarchy of the TCO 35
1 Introduction

Corpus-based linguistic studies have changed the field of linguistics in a striking manner – billions of indexed words provide an outstanding data basis for all kinds of analyses, ranging from machine translation to automatic speech recognition. However, there is one further aspect that is of noteworthy importance – novel linguistic theories can be crosschecked against real world data. Of course the reverse approach is possible as well – peculiar patterns that can be determined for those large sized data sets can be extracted and a theoretical framework can be developed in order to obtain an explanation for the considered phenomenon.

This thesis is meant to pursue this approach. My corpus based work during my internship at the Universitat Politècnica de Catalunya in Barcelona led to some highly interesting patterns that I want to explain in linguistic terms. But a careful evaluation is necessary in order to avoid being trapped by statistical artefacts. Therefore quantitative analyses play a major role in this thesis, as they allow unveiling underlying structures that are capable of describing the regarded problem in a much more effective fashion.

This thesis is divided into five parts – first, I am going to introduce the reader to the primal problem and its constituting data that motivated the present work. Moreover I will describe the applied methods that were necessary to perform quantitative analyses.

Section three will present an unexpected finding derived by the primal examinations, called the Strange Plot, that was produced by a certain parameter-setting, while the succeeding section will reveal the aspects causing that phenomenon. As a side effect, a surprising behaviour of the considered constructions will be detected.

Now that the usefulness of quantitative means is demonstrated, section five is heading towards an explanation for the primal task that is stated in the second section. Eventually the conclusion will summarise the present work and its results and will offer an outlook to future work.

2 Motivation and Methods

This section will cover important steps of my work at the Universitat Politècnica de Catalunya to clarify the basic elements of this thesis. I will describe the primal problem that we wanted to tackle and the means that were necessary in order to handle this problem. In addition I will provide an overview of the data that are the basis for further investigations.
2.1 The primal Problem

In our research group made up by Louise McNally, Berit Gehrke (both Universitat Pompeu Fabra), Gemma Boleda (Universitat Politècnica de Catalunya) and me, we investigated national-noun-constructions [1]. If you take, for example, the nation Spain and the noun wine, there are two constructions that are mainly used for combining them, namely Spanish wine and wine from Spain. I will denote these construction classes as AN-construction (for Adjective-Noun construction) and NpN-construction (for Noun-Preposition-Noun-construction). Moreover I want to denote the non-national noun as head-noun (in the given example that would be wine).

Although in most cases one construction could be substituted by the other, both classes of constructions exist and are frequently used. In order to explain the use of these kinds of constructions, we suggested context dependent reasons for selecting the proper construction.

Our research aimed at finding clues that trigger the generation of the respective construction based on corpus data.

2.2 The underlying Data

As the two constructions are very special types of noun-phrases we need a large-sized corpus for obtaining enough data, so that a various amount of analysis can be made to an useful extent. For this particular investigation the British National Corpus[2] (some more explanations in section 2.4.1) proved to be the right choice, since it represents a huge amount of text, namely 112.102.325 tokens, from a great variety of different sources.

Before extracting the noun phrases from the BNC we, of course, need some filter rules to establish a solid data basis. First of all we consider only nations whose nouns and adjectives occur at least 1.000 times. We also exclude the United Kingdom and the United States to avoid biasing towards these two states as they occur significantly more frequent than the other states. Furthermore these two nations might also have led to artefacts because we are using a British corpus and these two nations are used very broadly in the various newspaper articles amidst the BNC. The resulting 49 nations can be found in table 1.

Moreover one has to take care of nouns or possessive markers following the constructions as in these two examples:

(1) Education in Mexico City

(2) Chinese people’s feelings

Since we did not want to bias our data with possessive or multi-noun constructions, we needed to filter them out. Without a filter, a naïve query would match the
respective examples as the NpN-expression *Education in Mexico* and as the AN-expression *Chinese people*. Clearly, these expressions do not match the meanings of the real expressions – (1) is about Mexico City and not about Mexico and (2) is about the feelings of the Chinese and not about the people. Using a respective filter (that is to be found in section 2.4.1) the BNC contains 39,725 *AN-constructions* and 33,807 *NpN-constructions* where the 49 nations are involved. You can find the distribution of the data in table 1.

<table>
<thead>
<tr>
<th>Nation</th>
<th>AN</th>
<th>NpN</th>
<th>Nation</th>
<th>AN</th>
<th>NpN</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>2847</td>
<td>3802</td>
<td>Serbia</td>
<td>618</td>
<td>242</td>
</tr>
<tr>
<td>Germany</td>
<td>3400</td>
<td>2104</td>
<td>Denmark</td>
<td>515</td>
<td>325</td>
</tr>
<tr>
<td>Ireland</td>
<td>3367</td>
<td>1827</td>
<td>Mexico</td>
<td>406</td>
<td>382</td>
</tr>
<tr>
<td>Russia</td>
<td>3341</td>
<td>1182</td>
<td>Romania</td>
<td>396</td>
<td>368</td>
</tr>
<tr>
<td>India</td>
<td>2470</td>
<td>1411</td>
<td>Lebanon</td>
<td>336</td>
<td>397</td>
</tr>
<tr>
<td>China</td>
<td>2273</td>
<td>1380</td>
<td>Georgia</td>
<td>489</td>
<td>172</td>
</tr>
<tr>
<td>Japan</td>
<td>1701</td>
<td>1674</td>
<td>New Zealand</td>
<td>0</td>
<td>588</td>
</tr>
<tr>
<td>Spain</td>
<td>1917</td>
<td>1345</td>
<td>Croatia</td>
<td>295</td>
<td>293</td>
</tr>
<tr>
<td>Australia</td>
<td>1267</td>
<td>1391</td>
<td>Syria</td>
<td>333</td>
<td>230</td>
</tr>
<tr>
<td>Italy</td>
<td>1217</td>
<td>1273</td>
<td>Norway</td>
<td>171</td>
<td>385</td>
</tr>
<tr>
<td>Iraq</td>
<td>1261</td>
<td>990</td>
<td>Brazil</td>
<td>137</td>
<td>419</td>
</tr>
<tr>
<td>Poland</td>
<td>961</td>
<td>606</td>
<td>Cuba</td>
<td>301</td>
<td>238</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1214</td>
<td>306</td>
<td>Austria</td>
<td>176</td>
<td>352</td>
</tr>
<tr>
<td>Israel</td>
<td>524</td>
<td>963</td>
<td>Libya</td>
<td>306</td>
<td>219</td>
</tr>
<tr>
<td>Greece</td>
<td>1003</td>
<td>434</td>
<td>Pakistan</td>
<td>129</td>
<td>396</td>
</tr>
<tr>
<td>Turkey</td>
<td>953</td>
<td>458</td>
<td>Ukraine</td>
<td>277</td>
<td>239</td>
</tr>
<tr>
<td>Iran</td>
<td>635</td>
<td>607</td>
<td>Saudi Arabia</td>
<td>242</td>
<td>262</td>
</tr>
<tr>
<td>Switzerland</td>
<td>773</td>
<td>374</td>
<td>Belgium</td>
<td>206</td>
<td>292</td>
</tr>
<tr>
<td>Canada</td>
<td>379</td>
<td>701</td>
<td>Indonesia</td>
<td>247</td>
<td>209</td>
</tr>
<tr>
<td>South Africa</td>
<td>14</td>
<td>1028</td>
<td>Portugal</td>
<td>153</td>
<td>269</td>
</tr>
<tr>
<td>Egypt</td>
<td>319</td>
<td>660</td>
<td>Argentina</td>
<td>123</td>
<td>287</td>
</tr>
<tr>
<td>Sweden</td>
<td>546</td>
<td>424</td>
<td>Kenya</td>
<td>129</td>
<td>270</td>
</tr>
<tr>
<td>Vietnam</td>
<td>373</td>
<td>548</td>
<td>Jordan</td>
<td>140</td>
<td>249</td>
</tr>
<tr>
<td>Kuwait</td>
<td>236</td>
<td>650</td>
<td>Thailand</td>
<td>107</td>
<td>206</td>
</tr>
<tr>
<td>Hungary</td>
<td>502</td>
<td>380</td>
<td>Total</td>
<td>39725</td>
<td>33807</td>
</tr>
</tbody>
</table>

Table 1: The considered nations and the amount of adjective-noun and noun-preposition-noun constructions they are in.
2.3 Analysing that Data

For a further clean-up of the data, we introduced two more filters. To avoid a biasing due to infrequent constructions, we only considered head-nouns that occur, summed up over the two constructions, at least 25 times. Second, all head-nouns were excluded that were too nation-specific. This filter deals with nouns like reunification, subcontinent or wall. All of them are to be found mainly combined with one specific nation – for the given examples this would be Germany, India and China, respectively. In order to detect nation-specificity we used an entropy-based method that will be described in section 2.4.3 and ruled out every construction with an entropy lower than 3.

In the first analysis we simply listed the distribution over the two constructions for each of the 429 head-noun types that pass the filters and sorted them by their construction preference. That listing already gave the idea that abstract nouns prefer NpN-constructions and concrete objects especially person and institution denoting nouns prefer AN-constructions.

To pursue that intuition, we consulted the WordNet[3] based Top Concept Ontology (TCO)[4]. This ontology assigns features that are arranged in a hierarchy to each sense of a word. The three top-order concepts, subsuming all other features, are 1stOrderEntity, containing features describing concrete objects, 2ndOrderEntity, containing features describing events and situations and 3rdOrderEntity that is not divided into further features, but contains all nouns referring to abstract entities. Applying the TCO we could detect whether a given word refers to a concrete object or whether it denotes an abstract entity and compare the distribution of all head-nouns’ concepts for the two constructions.

More information about the TCO will be provided in section 2.4.2.

Indeed, our prior observation was supported and 1stOrderEntities are to be found more often in AN-constructions as opposed to 2ndOrderEntities and 3rdOrderEntities which prefer NpN-constructions.

Table 2 displays this behaviour. You can see the amount of tokens that can be found in the three main concepts and the Top concept that contains all words. Since the amount of AN-constructions and NpN-constructions differs, I use a relative frequency to calculate the ratio. This relative frequency is based on the Top concept of each construction. As you can see the number of 3rdOrderEntity tokens are almost equal for both constructions, but nevertheless the ratio indicates that there is a stronger preference for NpN-constructions as these constructions generally occur in a smaller amount compared to AN-constructions.

This table indicates that 1stOrderEntity head-nouns are expressed 43% more frequently in AN-constructions than in NpN-constructions. On the other hand NpN-constructions are used for 2ndOrderEntities and 3rdOrderEntities 70% and 19% more often, respectively, than AN-constructions.
Table 2: Distribution of the filtered tokens over the main concepts. Since the amount of AN and NpN constructions is not equal, a relative frequency concerning the Top-concept of a class was introduced in order to calculate the ratio.

Table 3 displays values for the same analysis but on a type basis. Except for 3rdOrderEntities with only few different types the differences are not as significant as in the token case. However, it is important to know the token distribution as well as the type distribution, since the token amount is an indicator of the frequency of usage, whereas the type amount indicates the degree of productivity. A high token count with only few types points to a class of words that are heavily used but not very productive – prepositions would be an extreme example as there is a huge token count adjoined by a tiny type count count, that implies no productivity at all, which is true, indeed.

Table 3: Distribution of filtered types over the main concepts.

The information given in the two tables above is summarised in figure 1. This plot is expanded by concepts whose hierarchy level is lower\(^1\) than 5 and that contain at least 10 different types. The intensity of one concept’s background colour indicates the hierarchy level – the darker the background the lower the level.

The y-axis shows the ratio of AN-construction vs. NpN-construction. A high ratio indicates a preference for AN-constructions and a low ratio indicates a preference

---

\(^1\)Please note that a low hierarchy level implies a more general concept. The Top-concept has hierarchy level 1.
for NpN-constructions. As the ratio is the result of a division, the axis has a log-scale such that a ratio of 0.5 has the same distance to 1.0 as 2.0, meaning that NpN-constructions occur twice as often as AN-constructions or the other way around, respectively.

The width of a bar points out the number of tokens that are subsumed by a given concept. Note that the width is scaled by the natural logarithm, indicating that twice the width means an about 7.4-fold amount of elements in a given concept.

Figure 1: This figure presents the distribution of the AN and NpN-constructions over the concepts given by the Top Concept Ontology. All types that occurred at least 25 times in the BNC and whose entropy is above 3.0 were considered.

Taking the plot into consideration the observation is further substantiated – 1stOrderEntities are above the 1.0 line and 2ndOrderEntities as well as 3rdOrderEntities are below that line. However, there are three remarkable exceptions:

**Part** containing types like border, part, territory, country, city, area, town, sea, empire, island, province, north, home.
Moreover it is evident that in almost every case the tendency is more extreme in the token case than in the type case. In the succeeding sections I will refer to this phenomenon as the **primal problem** and I will try to offer an explanation that offers reasons for this behaviour. But first of all I want to go into more detail concerning the applied methods.

### 2.4 A deeper Insight into the Methods

This part will describe the means that were necessary in order to establish the data basis in some more detail. First, I will describe how I accessed the *British National Corpus* and what kind of queries I used to extract the relevant noun-phrases. Second, the WordNet based Top Concept Ontology will be explained to some extent and third, I will show why entropy is a means of measuring specificity of head-nouns.

#### 2.4.1 Queries on the British National Corpus

For accessing and querying the *British National Corpus* (*BNC*) I used a tool from the *Corpus Work Bench* called *CQP*[5]. That tool allows an efficient search with complex queries and furthermore it is very useful for counting, sorting or grouping the findings as required. Since my work required lemma based queries, I used the XML version of the BNC, that eventually included a lemma feature. Of course I had to design two different queries for both classes of noun phrases:

**AN** Noun phrases of the AN-constructions have to start with an adjective derived from the 49 considered nations followed by an arbitrary noun. In addition this part must not be succeeded by another noun or possessive. The corresponding *cqp*-query looks like this:

```
/rnationa[ ] [pos="N.+"] [pos!="POS|N.+"]
```

2 All types are sorted by frequency.

3 A brief overview over the used part of speech tags: **NN0** – uncountable noun; **NN1** – singular noun; **NN2** – plural noun; **NP0** – proper noun; **POS** – possessive or genitive marker ‘s; **PRF** – of; **PRP** – prepositions except for of;
Please note that /rnationa[] is a macro that asks for one of the 49 national adjectives.

NpN Noun phrases of the NpN-constructions start with an arbitrary noun followed by a preposition and the national noun. Since the Netherlands ask for an article, the query also allows the definite article the in front of the national noun. Of course, the same filter as above is applied here. All that results in the following query:

[pos="N.+"] [pos="PR."] "the"? /rnationn[] [pos!="POS|N.+"]

please note that /rnationn[] is a macro as well that asks for one of the 49 national nouns.

2.4.2 WordNet and Top Concept Ontology

The EuroWordNet team tries to interlace the different wordnets that are available for different languages. They used interlingual Base Concepts to group clusters of words, no matter to which language they belong. Based on these Base Concepts they derived an ontology, the so called Top Concept ontology (TCO), that consists of 63 features that are divided into three main concepts:

1stOrderEntity Words that belong to this concept denote physical things such as government, people, troop, company, bank, border, part, army, state, official, party, leader, soldier, minister, society, woman.

2ndOrderEntity This concept contains words that denote events, states and properties like force, war, relation, tour, authority, art, history, life, time, trip, independence, policy, holiday, attack, agreement.

3rdOrderEntity All words denoting unobservable entities are part of this concept such as art, holiday, rule, year, experience, literature, proposal, frontier, style, example, tradition, right, demand, whole, plan.4

You can get a view of the complete ontology in appendix A. One should note that the TCO is no hierarchy in a strict sense, since words can be part of multiple sister nodes within the hierarchy tree, as concepts are regarded as features. The first layer below 1stOrderEntity consists of Form, Composition, Origin and Function and, of course, one word referring to a physical thing may have multiple of the mentioned features. Take a look at the assigned features for

4Note that the given examples are the most frequent head-nouns of each concept sorted by their frequency.
guitar – since it is a physical object, it belongs to the 1stOrderEntity concept. Now we need to consider its Form – it is of course an Object as opposed to Substance. As it is not Natural its Origin is Artificial, what makes good sense too. Its Function is correctly classified as Instrument. However, as the provided examples in the listing above document, the classification unfortunately is not always as precise as for the guitar case. It is disputable, at least, whether words like border, part or party refer to physical things – especially in a national context. But since the TCO covers a broad amount of words, one can assume that tendencies can be derived correctly – in particular with regard to the type counting case.

To examine other aspects of the TCO, it is necessary to see how words with multiple senses are handled. It can cope with that by working with the sense indices that WordNet assigns to each of the senses of a given word. For that reason TCO does not list features of words, but it lists features of senses.

How do I determine the correct sense for a word, say party in the expression party in China then? One possibility would be to consider the most frequent sense only. It is possible to access this information, since the senses are ranked according to their frequency by WordNet. However, just considering the first sense might decrease precision a lot. Another alternative would be a sense disambiguation of the BNC before extracting the noun phrases in order to use context information for identifying the correct sense more reliably. But due to time restrictions I restricted myself to an approximation – I assumed that the most frequent sense occurs twice as often as the second most one and the second ranked sense occurs twice as often as the sense on place three and so forth. I derived a formula that returns a relative frequency given a rank $r$ and an absolute number of senses of a word $n$:

$$relFreq(r, n) = \frac{k^n(1 - k)}{k^r(1 - k^n)}$$

This formula also takes care of normalisation constraints such that

$$\sum_{r=1}^{n} relFreq(r, n) = 1$$

$k$ is a constant that represents the ratio of the value of a rank and the value of its succeeding rank. In my case it has the value 2.

For my analysis of the distribution of the head-noun constructions over the different concepts, the value of the senses of a polysemous word were weighted according to their relative frequency as described above. I want to illustrate the formula on an example – the noun plan has, according to WordNet, three senses:

**sense 1** “a series of steps to be carried out or goals to be accomplished”
concepts  Top, 2ndOrderEntity, SituationComponent, Cause, Purpose, 3rdOrderEntity

sense 2  “an arrangement scheme”

concepts  Top, 2ndOrderEntity, SituationComponent, Experience, Mental, SituationType, Static

sense 3  “scale drawing of a structure”

concepts  Top, 1stOrderEntity, Form, Object, Function, Representation, ImageRepresentation, Origin, Artifact

plan occurs 38 times in an AN-construction in the BNC. Due to the formula above, the token number would be split up into the following partitions

sense 1  21.714

sense 2  10.857

sense 3  5.429

When plan gets inserted into an AN listing where token counts for each concept are considered, Cause, for example, would be increased by 21.714, 1stOrderEntity by 5.429, 2ndOrderEntity by 32.571 and Top by 38. If that listing was a type count all these values would be divided by the frequency, this is 38, before inserting them into the respective concept count column.

This means that all the numbers presented in table 2 and 3 are no natural numbers but fractions that were rounded. Of course this does not account for the Top concept, since every sense has a Top concept.

2.4.3 Entropy as a Means of measuring Specificity

As already said there are some head-noun types that have many tokens in the corpus but do only occur with one or very few nations. Since these fixed nation-noun constructions may bias the analysis in a significant manner, I needed to find a means that was able to tell well distributed head-nouns apart from specialised ones.

Fortunately information theory provides us with a measure that is able to quantify the uniformity of distribution – it is called entropy [6].

If there is no distribution at all, this is, there is only one outcome associated with the regarded event, the entropy would be 0. The entropy has a maximal value for a given amount of outcomes, if the outcomes are uniformly distributed.

At this, it is important to know that a higher amount of uniformly distributed
outcomes results in a greater entropy. The formula for calculating the entropy of $X$ with $n$ outcomes is as follows:

$$H(X) = -\sum_{i} p(x_i) \log_2 p(x_i)$$

where $X$ is the probability distribution and $p(x_i)$ denotes the probability of the $i$-th outcome.

I want to provide two examples in order to clarify how it works:

**reunification** In AN-constructions this word occurs only with two nations in the BNC – 49 times with *Germany* and once with *Romania*. The corresponding probability distribution looks like this:

$$\begin{align*}
\Pr(X = \text{Germany}) &= 0.98 \\
\Pr(X = \text{Romania}) &= 0.02
\end{align*}$$

Inserting this values into the formula above returns $\approx 0.141$. A very low value which makes good sense, as the distribution is very one-sided.

**plan** This word occurs in 19 NpN-constructions and the probability distribution is as follows:

$$\begin{align*}
\Pr(X = \text{Germany}) &= 0.158 \\
\Pr(X = \text{Croatia}) &= 0.105 \\
\Pr(X = \text{Poland}) &= 0.105 \\
\Pr(X = \text{France}) &= 0.105 \\
\Pr(X = \text{Japan}) &= 0.105 \\
\Pr(X = \text{Cuba}) &= 0.105 \\
\Pr(X = \text{Sweden}) &= 0.053 \\
\Pr(X = \text{Russia}) &= 0.053 \\
\Pr(X = \text{Italy}) &= 0.053 \\
\Pr(X = \text{Argentina}) &= 0.053 \\
\Pr(X = \text{Turkey}) &= 0.053 \\
\Pr(X = \text{Iraq}) &= 0.053
\end{align*}$$

This almost uniform distribution results in an entropy of $\approx 3.471$. Note that a perfect uniform distribution over the 12 nations would have led to an entropy of $- \log_2 \frac{1}{12} \approx 3.585$
Figure 2: Three sample head-nouns annotated with their entropy and their nation distribution.

Each head-noun has two entropy values – one for its nation distribution in AN-constructions and the other one for NpN-constructions. Figure 2 displays with which distributions the different entropies are associated. Each section of a bar represents one nation’s portion. federation as well as south are mainly to be found accompanied by one specific nation (Russia and France respectively) whereas army is quite well distributed involving 37 nations. After manually exploring the data I decided that 3 would be a good lower bound for the entropy to guarantee a good distribution.

3 The Strange Plot

Before tackling the big question of what leads to the dissimilar behaviour of 1stOrderEntities and 2ndOrderEntities for the two constructions, I want to demonstrate the importance of profound quantitative analysis by applying it on a more specific phenomenon that I encountered. This section will describe the mentioned phenomenon called the Strange Plot which is derived from the primal problem.
3.1 Origin of the Strange Plot

During experimentation with the parameters, i.e. entropy, token or type counts and minimal token frequency, I discovered an unpredicted behaviour, when considering the usual minimal token frequency of 25, no entropy filter and type counts. This set-up results in an almost uniform distribution where every concept has a ratio very close to 1. Since constructions with an entropy greater than 3 behave as expected\(^5\) one must conclude that constructions with an entropy lower than 3 have to behave in a mirrored manner. The corresponding plot is called the *Strange Plot* and can be found in Figure 3.

![Ratio Distribution over Concepts](image)

Figure 3: This figure displays the *Strange Plot*. It shows the distribution of the AN vs NpN ratios over the given concepts for three different entropy bounds.

The axes and the blue bars are the same as in figure 1, as type counts are considered. The green bars follow from not filtering the data by entropy which results in a distribution close to 1 in all cases. They are the result of the two distinct

\(^5\)You can find the corresponding data in the blue bars in figure 1.
data sets consisting of ratios from head-nouns with an entropy greater or equal to 3 (blue bars) and ratios from head-nouns with an entropy smaller than 3 (yellow bars).

3.2 The Strange Plot from another Point of View

In order to exemplify how the data has to be distributed to result in this phenomenal observation, I want to provide a sample set of head-nouns for a concept class that only contains few members. I chose words whose most probable sense lies within the concept of *Representation* which is a sub-concept of *Function* and *1stOrderEntity*. (See Section A for the hierarchy of concepts in the Top Concept Ontology)

Please bear in mind that each head-noun has one entropy for each construction class. The word *call*, for example, has an entropy smaller than 3 in AN-constructions but an entropy greater than 3 in NpN-constructions.

**Entropy < 3**

- **AN-constructions** call, crown, entry, figure, letter, map, picture, qualifier, return, sculpture, share, software, target, title – 14 types
- **NpN-constructions** crown, currency, entry, fund, letter, magazine, medium, picture, qualifier, radio, sculpture, share, sign, software, television, tv – 16 types

**Entropy >= 3**

- **AN-constructions** book, currency, fund, loan, magazine, medium, money, name, newspaper, radio, record, report, sign, television, term, translation, tv, word – 18 types
- **NpN-constructions** book, call, figure, loan, map, money, name, newspaper, record, report, return, target, term, title, word – 15 types

This listing should visualise the phenomenon quite well. For an entropy bigger than three the ratio is as expected – within the AN-constructions there are 18 types contrasted to 15 types of NpN-constructions – a preference for AN-constructions. For the small entropy case the behaviour is mirrored – there are 16 NpN-constructions and only 14 AN-constructions. If you do not consider any entropy at all the ratio is almost balanced – 32 AN-constructions and 31 NpN-constructions.

Since all these numbers are quite close to each other, one could suggest that this behaviour is a random pattern that incidentally meets the described conditions. But as this pattern occurs within all concepts this assumption is very unlikely.
3.3 Finding Hypotheses explaining the Strange Plot

Before finding an explanation for the *Strange Plot*, I want to rephrase the problem in order to ensure that its core message is understood. We could see that nation-unspecific head-nouns, this are head-nouns with a high entropy, behaved as expected, whereas, on the other side, nation-secific nouns behaved in a very dissimilar manner. This means that nation-specific head-nouns prefer NpN-constructions, if they belong to an 1stOrderEntity-concept and they prefer AN-constructions if they belong to abstract concepts.

In order to find an explanation for this phenomenon there are two main paths one can follow. Either one looks for statistical artefacts that may be caused by an unfortunate combination of parameters or, if that approach should fail, one looks for a linguistic theory that accounts for the findings.

Of course the primal problem remains, if the reasons that constitute the *Strange Plot* are found. The primal problem will be discussed in section 5 then. In the next section I will examine in how far the application of quantitative methods will yield useful results in order to explain the *Strange Plot*.

4 Advance towards the Core of the Strange Plot

After describing the *Strange Plot*, I now want to extract the core features that are constituting this phenomenon by applying quantitative analysis.

4.1 Type based Analysis

Examining the sample set given in section 3.2 carefully, a striking observation can be made – almost every word occurs twice, once in the AN-construction listing and once in the NpN-construction listing. That is what motivated the following contingency table:

<table>
<thead>
<tr>
<th>AN-const.</th>
<th>entropy &lt; 3</th>
<th>entropy &gt; 3</th>
<th>NpN-constructions</th>
<th>entropy &lt; 3</th>
<th>entropy &gt; 3</th>
<th>does not occur</th>
<th>NpN-constructions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>115</td>
<td>131</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td>257</td>
</tr>
<tr>
<td></td>
<td>112</td>
<td>170</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>287</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>11</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>228</td>
<td>312</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>556</td>
</tr>
</tbody>
</table>

This table takes every occurring type that passes the amount filter into consideration and classifies it according to its AN-construction entropy and NpN-construction entropy. If it does not occur in one of the constructions it is inserted into the respective does not occur-field. One should have noted that the table
provides very important information – only 16 head-noun types that occur in AN-constructions do not occur in NpN-constructions and there are only 12 head-noun types that exclusively occur in NpN-constructions. This is of crucial importance in order to explain the Strange Plot. As almost every word is listed in the AN-construction class and the NpN-construction class, the ratio of AN-construction and NpN-construction of nearly 1.0 is a necessary result.

To put it in other words, the Strange Plot is constituted as follows:
for each concept there are two sets of words – one for AN-constructions and the other for NpN-constructions. Virtually every head-noun of the filtered data basis is part of both sets. Since only type-counts are considered the size of the two sets is almost the same for every concept. That is what causes the ratio to be 1, unless there is no filtering by entropy.

Splitting this bulk of head-nouns into two distinct parts has to result in a mirrored behaviour of those two parts to guarantee that the combined ratio is close to 1. As we already have seen, the part where nation-specific head-nouns are ruled out, has an explicit distribution characteristic. Therefore the other part’s ratios need to have the corresponding symmetrical values.

4.2 Gained Insights

This surprisingly simple explanation comprises one very important message – the data distribution that causes the Strange Plot is a mere statistical matter. Although there might be a different behaviour of nation-specific head-nouns compared to non-nation-specific ones, that behaviour cannot be grasped with the current analysis. But with the present knowledge of the data characteristics, one should be able to establish a parameter setting that avoids the shortcomings of the prior approach.

For nation-specific nouns I applied a much more restrictive filtering. For the strange plot it sufficed that a head-noun occurred at least 25 times, summed up over both constructions. Therefore it often happened that nouns were considered as nation-specific for one construction due to a low frequency count and a resulting low entropy. For example the word change – it occurs 50 times in the BNC in an NpN-construction, with an entropy of 3.402. However, it occurs only once in an AN-construction and is treated as nation-specific in the AN-constructions although no information about specificity can be made, because there is not enough information. Intuition, though, says that change is no nation-specific noun for AN-constructions.

Taking this into consideration, a reasonable filter for extracting nation-specific nouns would consider head-nouns that occur at least 25 times in one construction and have an entropy smaller than 2 in this particular construction. The filter for non-nation-specific nouns was just marginally changed, instead of asking for
25 minimal occurrences, summed up over both constructions, it needs to have at least 25 token counts in the respective construction in order to be inserted. In other words, if there was head-noun whose AN-construction occurred 10 times and NpN-constructions 30 times, the prior filter would consider both constructions as the sum is 40. The current filter would exclude the AN-construction and would only consider the NpN-expression for further analysis.

As already mentioned, the nation-specificity filter is very restrictive and therefore only very few different types exists – 37 for AN-constructions and 7 for NpN-constructions. That is why I decided to consider token counts for the new analysis only. Figure 4 presents the data distribution given the presented parameters for telling specific and non-specific head-nouns apart.

Figure 4: This figure displays the altered Strange Plot. It shows the distribution of the AN vs NpN token ratios over the given concepts for nation-specific head-nouns with a low entropy and non-nation-specific head-nouns with a greater entropy.

The plot shows every concept that contains at least 500 tokens, summed up over the two noun classes, and the respective ratio of nation-specific and non-specific
head-nouns. There are some bars like the yellow bar for Building that go beyond the scope of the plot. This is due to a one-sided data distribution. Since there is no nation-specific Building head-noun in an NpN-construction, the bar reaches to infinity. This plot still seems to conserve some features from the Strange Plot. It is even more remarkable, if you consider that the Strange Plot set-up for token counts did not work out – it was similar to figure 1. The new filters, however, seem to transfer some characteristic behaviours of the Strange Plot to token count distribution. This might be explained with the hypothesis that token counts could be effectively considered as type counts due to the small amount of different types. However, because of the small amount of nation-specific nouns, a qualitative analysis seems to be appropriate now.

4.3 Analysing the new Approach

Although the amount of detected nation-specific head-nouns is dangerously small for a profound analysis, one, nevertheless, can infer some important insights. But before working with the graph, it might be useful to take a look at a complete listing of the nation-specific head-nouns:

**nation-specific head-nouns in AN-constructions** summer, cypriot, spy, wall, house, conquest, orphan, bath, subcontinent, sea, tribe, investment, referendum, mark, aborigine, unity, measles, franc, assembly, federation, chancellor, time, unification, mainland, reunification, tragedy, bread, building, peseta, rugby, roulette, dollar, opera, rearmament, succession, mythology, confederation

**nation-specific head-nouns in NpN-constructions** gulf, hostage, god, crown, port, bird, south

Regarding this collection of words, one can infer that this listing contains mainly words that are, indeed, specific to a nation. The filter seems to work well. Nevertheless some artefacts can be observed like Georgian Building or Port of Spain. This is what motivated a classification into the following three categories:

**sampling error** spy (Russian and German spies), house (Georgian house), orphan (Romanian and Russian orphans), tribe (Indian tribe), investment (Japanese investments), referendum (Danish referendum), assembly (only Serbian and Croatian Assembly), Irish Time, building (Georgian building), rugby (mainly Irish and some French Rugby and others), rearmament (German rearmament), succession (War of the Spanish Succession), hostage (hostages in Lebanon), crown (Crown of France), bird (Birds of Australia [a book]), south (south of France)

---

6 the bracketed expressions name the specific nation(s) that refer to the head-noun.
geographic term subcontinent (Indian Subcontinent), sea (Irish Sea), federation (Russian Federation), mainland, (Chinese mainland), confederation (Swiss and German Confederation), gulf (Gulf of Mexico), port (Port of Spain)

non-geographic multi-word-expression summer (Indian Summer), cypriot (Greek and Turkish Cypriot), wall (Chinese Wall), conquest (primarily Spanish and Russian Conquest), bath (Turkish Bath), mark (German Mark), aborigine (Australian Aborigine), unity (German and some Irish unity), measles (German Measles), franc (French Franc), chancellor (German chancellor), unification (German unification), reunification (German reunification), tragedy (Greek Tragedy), bread (French bread), peseta (Spanish Peseta), roulette (Russian Roulette), dollar (Canadian and Australian Dollar), opera (Italian Opera), mythology (Greek Mythology), god (God of Israel)

What were the criteria for classifications? Sampling errors refer to pretended nation-specific head-nouns that occur due to extensive use in only few texts within the corpus – words that are likely to occur with other nations as well in a different context.

geographic terms denote expressions that can be found in an atlas and non-geographic multi-word-expression refer to terms where there is a fixed expression which is exemplified in either a capitalised head-noun and/or an article in the wikipedia. Of course this rules can be considered as arbitrary or, at least, fuzzy, if you consider for example German rearmament vs. German unity. But you can swap the borderline cases as you wish – except for God of Israel there is no NpN-construction that could enter into the last item of the list. And even God of Israel is a special case, as it only occurs in citations from the Bible and therefore could not be considered as modern English. Rejecting God and the geographical terms, there is no valid NpN-construction and therefore considerations concerning figure 4 can be neglected, as it is of no use to examine the ratios of NpN vs. AN-constructions anymore.

How does it come then that only AN-constructions qualify for that kind of lexicalisation? Is there anything that prevents NpN-constructions from forming lexical entries? Of course the list above is not complete and surely there are fixed NpN-expressions as well. The wikipedia query on German chancellor redirects to the entry of Chancellor of Germany, for example. However, the last item of the mentioned list contains many entries that are very hard to transform into an NpN, if it is possible at all. What about Turkish Bath, for instance? It is no Bath in Turkey, nor is it a Bath from Turkey and a Bath of Turkey is inappropriate as well. It is rather a Bath whose style origins from Turkey. A relation that is very difficult to grasp with a preposition.

An extensive use for AN-constructions for fixed expressions can therefore be explained by the fact that complex relations can be expressed by a short construction,
since the relation between the nation and the noun is expected to be made in the head of the listener and not in the words themselves. The presupposition, that the addressee knows the relation, can be legitimated, as we speak of fixed expressions that can be considered as common knowledge.

5 Advance towards the Core of the primal Problem

Now, that the mystery that constitutes the Strange Plot is unwrapped and we gained some interesting insights, the cause for the primal problem is still left unclear. What is the reason for non-nation-specific head-nouns to prefer AN-constructions when the head-noun is a 1stOrderEntity and why do they prefer NpN-constructions otherwise? To answer these questions, I, once more, will regard the data from different perspectives. The interplay of type entropy and type frequency will play a major role in these considerations.

5.1 Entropy Distribution

This part aims at unveiling the entropy distribution for both constructions over the three main concepts 1stOrderEntity, 2ndOrderEntity and 3rdOrderEntity. For obtaining a first impression take a look at the box plots in figure 5. Please note, that the considered data is unfiltered since the unfiltered data-set conveys more valuable information concerning our problem.

The red box which represents the entropy distribution of AN-constructions has a greater median than the green box representing the NpN-construction’s distribution for 1stOrderEntities, whereas for 2nd and 3rdOrderEntities the ratio behaves in a contrary manner. Moreover it is remarkable that more than half of the 1stOrderEntropy NpN-constructions have an entropy value of 0. This can be explained by Zipf’s Law[7] that states that there are only few types with many tokens and many types with few tokens, and small token amounts lead to small entropies. To get a better idea of how exactly the entropies are distributed I provide a histogram in Figure 6.

As usual the red colour represents AN-constructions and NpN-constructions are coloured green. Where one construction outranges the other, the intermediate area is coloured according to the surpassing construction. Moreover a bar-based histogram was inserted in the background to provide an impression of the amount of types that are considered. The bars do not show the distribution of both constructions but the distribution of only that construction with the highest peak. This figure provides even better evidence for the assumption that there is a difference of the entropy distributions for every concept. Unfortunately the distribution
Figure 5: This box plot shows the distribution of the entropy over the three main concepts where the red boxes represent the AN-constructions and the green boxes represent the NpN-construction.

Figure 6: This graph shows a histogram for the three main concepts. Red lines represent AN-constructions and green lines NpN-constructions. The bars in the background show the frequency of the entropy partitions.
does not correspond to a standard normal distribution so that no t-Test can be applied. The given distribution is caused by Zipf’s Law as well. Since a small token amount presumably leads to a small entropy, most entropies are rather small. This also causes the valley between 0 and 1 in the three plots – there are many head-nouns occurring only with one nation resulting in an entropy of 0. To obtain an entropy slightly greater than 0 there needs to be a head-noun that often occurs with one nation and rarely occurs with other nations. This is much more unlikely than a head-noun that occurs few times with few nations which results in an entropy close to 1.

5.2 Frequency Distribution

Before evaluating whether there is a dependency of entropy and frequency, I first want to display the frequency distribution. Therefore I will display the box plot and histogram once more for this feature. Note that the frequency counts are on a logarithmic scale. With regard to the box plot, one can observe a slightly different behaviour for the three concepts as compared to the entropy distribution.

First of all, it is interesting to see that the median for all three concepts is 2, second, there have been some changes concerning the ratios – while ratio for 1stOrderEntities stays the same, it is hard to determine a ratio for 2ndOrderEntities. Compared to the entropy box plot the ratio of 3rdOrderEntities is even conversed. Considering the histogram in figure 8, the observed behaviour persists. These histograms as well behave according to Zipf’s Law - as already indicated in the preceding section, the majority of head-nouns has a small token count. The valleys between 0 and 0.5 are a consequence of the logarithmic scale of the x-axis. The logarithm of 1 is 0; the logarithm of 2 to base 10 is approximately 0.3. As only natural numbers are considered, there is no token amount between 0 and 0.3 which results in the observable valley.

Nevertheless it is interesting to note that, at least for 1st and 3rdOrderEntities, there is a tendency for high-frequent head-nouns to prefer AN-construction, while low-frequent head-nouns tend to prefer NpN-constructions.

5.3 Correlation of Entropy and Frequency

One shortcoming of an entropy based analysis is the fact that the resulting entropy value depends on the frequency of the considered head-noun – there always exists an upper limit for the entropy that can be determined by calculating $\log_2(n)$ where $n$ is the amount of tokens for the respective head-noun. Of course, this upper limit can only be approached, if the $n$ tokens are partitioned in such a way that there is only one occurrence per nation. Since there is a maximum of 49 nations the
Figure 7: This box plot shows the distribution of the concepts type counts. The red boxes represent the AN-constructions and the green boxes represent the NpN-construction.

Figure 8: This graph shows a histogram for the three main concepts. Red lines represent AN-constructions and green lines NpN-constructions. The bars in the background show the logarithm to base ten from the frequency counts of the respective types.
absolute upper limit for the entropy is $\log_2(49) \approx 5.615$.
However, the preceding two sections already indicated that there is an, at least slightly, different behaviour of entropy and frequency distribution. Now I want to provide some further evidence in order to demonstrate, whether the entropy values are a mere side-effect of frequency counts or if they contain valuable independent information.
That is why I provide a plot that displays each head-noun as two dots whose positions are dependent on the frequency and the entropy for both constructions. AN-constructions appear as red circles and NpN-constructions as green squares. Moreover I added a line that marks the mentioned upper bound.

![Figure 9](image)

Figure 9: Each head-noun is represented by two dots whose position is dependent on frequency and entropy of the two constructions. As usual AN-constructions are displayed red and NpN-constructions are coloured green.

Since there exist some frequency-entropy constellations where multiple head-nouns occur, I added some noise to each data point’s position, so that one can estimate...
the amount of points at named locations. However, for low frequent types there are so many types that the red points are superimposing the green points under-
neath.
The figure already indicates that there is some correlation but, nevertheless, the point distribution is too fuzzy to claim for an explicit interdependence.
As the separation of this point cloud into the three main-concepts might supply some more insights, take a look at figure 10.

Figure 10: Each head-noun is represented by two dots whose positions depend on frequency and entropy of the two constructions. As usual AN-constructions are displayed as red circles and NpN-constructions are coloured green.

Throughout all three plots an observation can be made which is visualised by the green and the red line\(^7\) – AN-constructions with a high token count tend to have a lower entropy, whereas NpN-constructions are closer to the upper limit line, implying a greater entropy for the same token frequency.
This observation conveys the important message that entropy should not be neglected due to a putative correlation but may hold important independent information instead. A discussion of that matter will be provided in section 5.5.

5.4 Further Reflections concerning Entropy
As already explained, it holds that entropy is bound to sample size to some extent. However, if we assume that we know the real entropy for a given head-noun in a given construction due to an exploration of an infinitely sized corpus, we might be

\(^7\)The lines were created by applying locally-weighted polynomial regression[8] – a smoothing technique mainly used for visualisation
able to create a model to interpolate the real entropy, given a sample entropy and the sample size. The real entropy of a type that would be obtained by consulting an infinitely sized corpus, is called *population entropy* as opposed to the actually measured *sample entropy*.

For the following disquisition I suggest three population entropies that are derived from three frequent head-nouns with differing entropies. You can find the three head-nouns and their nation distribution in figure 2. Based on the relative nation-count for each head-noun, I derived several (50) random samples for a given size (10, 15, 25, 40, 66, 100, 150, 250, 420, 660 and 1000). I calculated the sample entropy for each of these samples and plotted them in figure 11. Regarding the red samples that are derived from the head-noun *army*

![Figure 11: Sample entropy distribution for three imaginary population entropies based on the nation distribution of the head-nouns army (red), south (blue) and federation (green).](image)
with an entropy of 4.485, one can see that the sample entropies are increasing on average when the sample size increases, while the standard deviation, which is indicated by the dotted lines, decreases, as it does in the other two examples as well. At a sample size of 1000 the mean sample entropy is about 4.45.

The blue samples are derived from the nation-distribution of *south* that has an entropy of 1.903. The average sample entropy rises as well, but in a flatter fashion and the standard deviation is the greatest of all three examples. The mean sample entropy for a sample size of 1000 is about 1.88.

Eventually, the green samples’ average sample entropies oscillate at the beginning and stay at a level of about 0.76 for greater sample sizes. That sample is derived by the *federation’s* nation distribution and has an entropy of 0.791.

As expected the average sample entropy is below the actual population entropy. Therefore one can conclude that most of the examined entropies that were extracted from the BNC must have a higher value. But establishing a robust mathematical model which achieves a good approximation for a given sample entropy and corresponding sample size is a very difficult task, since for sample sizes smaller than 50 the variation of the sample entropies is too big to establish a precise prediction of the intrinsic population entropy.

As most of the head-nouns have a token frequency significantly below 50 such a model would cause many dubious results. Even the filtered head-nouns have a median frequency clearly below 50.

But since the model would treat AN and NpN-constructions in the same manner, there would probably be no further information gain that would support finding an explanation for their different behaviours over the different concepts. However, the given analyses could be more precise, if they relied on population entropies.

### 5.5 Final Considerations

Collecting all the information of the current section, there are some important hints concerning entropy and token frequency that lead to a sound explanation of the considered AN and NpN-construction distribution. First, there was the assumption that NpN-constructions have a greater entropy in *2nd* and *3rdOrder-Entities*, as opposed to *1stOrderEntities* where AN-constructions have a slightly greater entropy, as shown in figure 5 and 6.

This can be explained by a higher flexibility of NpN-constructions for creating new terms, when applied to *2nd* and *3rdOrderEntities*. AN-constructions, on the other hand, seem to be rather restricted concerning the creation of new words as the lower entropy for high frequent head-nouns indicates for these two concepts.

Figure 8 supports this thesis as it demonstrates – especially for *1st* and *3rdOrder-Entities* – that there is a high amount of low frequent head-nouns from NpN-constructions, whereas established, high frequent nation noun combinations are
rather to be found for AN-constructions. This is also supported by figure 10.

Now, that we know that NpN-constructions are more likely to create new nation-
noun combinations, we gained some more important insight. But how can we
explain the differing behaviour over the different concepts?

Figure 8 can offer an explanation – as already stated, head-nouns with a low
frequency count, including many novel nation-noun combinations, are rather to
be found in NpN-constructions while AN-constructions occur more often in es-
tablished constructions. This seems to be different for 2ndOrderEntities – AN-
constructions are not able to establish more head-nouns with high frequency than
NpN-constructions. In fact they behave very similar for that concept.

More support for this matter can be gained by consulting table 2 and 3. One can
derive the average token count for each construction and concept which results in
table 4.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Average AN tokens</th>
<th>Average NpN tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>67.7</td>
<td>51.9</td>
</tr>
<tr>
<td>1stOrderEntity</td>
<td>71.6</td>
<td>48.0</td>
</tr>
<tr>
<td>2ndOrderEntity</td>
<td>60.3</td>
<td>57.1</td>
</tr>
<tr>
<td>3rdOrderEntity</td>
<td>74.1</td>
<td>48.0</td>
</tr>
</tbody>
</table>

Table 4: Average token count for each concept over the two constructions.

This table\(^8\) confirms the proposed hypothesis – in general the average token size is
greater for AN-constructions. For 2ndOrderEntities this still holds, but compared
to its average count, there are remarkably greater NpN-construction token counts,
whereas AN-constructions have a lower count than in the other concepts.

Obviously, it is harder to establish AN-constructions for 2ndOrderEntities which
results in a higher use of NpN-constructions.

Now it seems that this finding is contradictory to the findings that were made
while examining the Strange Plot where I stated that mainly AN-constructions
are able to create fixed nation-noun combinations. However, there are some major
differences – first of all the Strange Plot findings were based on nation-specific
head-nouns. As we now consider all types of nouns, the nation-specific data set
only presents a small minority of the regarded data. Moreover, I do not want to
say that there are no fixed AN 2ndOrderEntity expressions at all. What I want to
express is the hypothesis that stable AN-constructions are more infrequent to be
found in 2ndOrderEntities as compared to other concepts.

Intuitively this makes good sense – if you have a 1stOrderEntity, say a person, there
is only one meaningful relation to a nation – the origin relation which usually would

\(^{8}\)Please note that this table is based on the filtered data set from table 2 and 3. However, the
tendency stays the same if there would be no filtering.
be expressed by the prepositions from or of. Because there is no ambiguity, it is common to use an AN-construction as it is shorter and more convenient to use. 2ndOrderEntities, however, often refer to nouns that can be related to a nation in multiple ways. If you encounter an expression like Serbian agreement you cannot be sure about the relation of Serbia and agreement. It could be an agreement made in Serbia, an agreement with Serbia or an agreement about Serbia. Considering, for example, Greek artist there would be no such ambiguity in the general case. Nevertheless, there is still contradictory evidence for 3rdOrderEntities. According to token counts and figure 8 there are more head-nouns with high frequency belonging to AN-constructions, but still, as displayed in figure 1, there is a preference for NpN-constructions. But as this concept is rather small containing only 21 AN and 32 NpN-constructions, it is quite vulnerable to outliers. Indeed, government and art make up almost one third of the token mass from the AN-constructions. If these two words are excluded, 3rdOrderEntities quantitatively behave much more like 2ndOrderEntities.

6 Conclusion and Outlook

To summarise the present work, it can be stated that the applied quantitative analyses led to new insights, making it possible to describe and explain the considered phenomena by its core features – the phenomenon of the Strange Plot could be reduced to a statistical artefact. However, the knowledge about the cause for this artefact was used in order to rephrase the rules for extracting nation-specific and non-nation-specific nouns. The novel formulation eventually led to the surprising finding that AN-constructions are much more likely to form multi-word-expressions, as compared to NpN-construction, due to the lack of necessity of describing the relation explicitly, what explains nation-specific nouns’ preference for AN-constructions.

The last part of this work was a quite lengthy analysis of the interplay of frequency and entropy over the different concepts and constructions. This work, however, paid off, as some of the analyses led to insights that were very useful for finding features that allowed to explain the problem at hand.

The examination of the two considered phenomena led to the conclusion that AN-constructions are used more often for stable expressions whereas NpN-constructions are rather used to introduce novel expressions. Moreover it is important to note that, in addition to the described effect, there is a preference for AN-constructions to relate a nation with an physical object, since for these entities the relation is easier to get – usually the nation denotes the origin of the object. Opposed to this behaviour, NpN-constructions are preferred for expressing a relation between a nation and an abstract noun, since this relation is harder to grasp without a
preposition. This findings also support the initial suggestions that were made during my work at the UPC and are now justified by quantitative results. Future work on this topic could include a more profound analysis of more delicate aspects of the problem such as finding the reasons that result in the the outliers Place, Part and Modal. One preliminary hypothesis would suggest that the mentioned caveats of the TCO might be a cause. However, there are more facets that could be considered. For example word sense disambiguation could be applied or a mathematical model for obtaining the population entropy could be developed in order to increase the accuracy of the findings. Taking all this into consideration, I successfully could apply quantitative methods to receive a solid basis for sound qualitative explanations.

7 Acknowledgements

First, I would like to thank Stefan Evert who proved to be a great supervisor, giving many important advices that enabled me to develop this thesis as it is – no matter how critical the situation has been. Moreover I want to thank Louise McNally for her helpful comments, that were crucial for increasing the quality of this work. Furthermore I would like to thank Gemma Boleda who guided me in a wonderful manner through my internship at the UPC, arousing my interest in this kind of corpus based linguistics mashed with theoretical aspects. Eventually, I want to thank Paula Vieweg and Tobias Springenberg, who made important nodes on the draft that helped a lot to improve the quality of the present thesis as well.
References


[4] Javier Alvez; Jordi Atserias; Jordi Carrera; Salvador Climent; Egoitz Laparra; Antoni Oliver and German Rigau, Complete and Consistent Annotation of WordNet using the Top Concept Ontology, Sixth International Language Resources and Evaluation (2008).


A Feature Hierarchy of the TCO

Top
- 1stOrderEntity
  - Form
    - Object
    - Substance
      - Gas
      - Liquid
      - Solid
  - Composition
    - Group
    - Part
  - Origin
    - Artifact
    - Natural
      - Living
        - Animal
        - Creature
        - Human
        - Plant
  - Function
    - Building
    - Comestible
    - Container
    - Covering
    - Furniture
    - Garment
    - Instrument
    - Occupation
    - Place
  - Representation
    - ImageRepresentation
    - LanguageRepresentation
    - MoneyRepresentation
  - Software
  - Vehicle
2ndOrderEntity
  SituationType
    Dynamic
      BoundedEvent
      UnboundedEvent
    Static
      Property
      Relation
  SituationComponent
    Cause
      Agentive
      Phenomenal
      Stimulating
    Communication
    Condition
    Existence
    Experience
    Location
    Manner
    Mental
    Modal
    Physical
    Possession
    Purpose
    Quantity
    Social
    Time
    Usage

3rdOrderEntity
Eidesstattliche Erklärung

Declaration of Authorship

Hiermit erkläre ich, die vorliegende Studienarbeit zur Erlangung des Bachelors of Science in Cognitive Science selbständig verfasst und keine anderen Quellen und Hilfsmittel als die angegebenen verwendet zu haben.

I hereby confirm that I have written the present thesis for obtainment of the Bachelor’s degree of Science in Cognitive Science independently, and I have not made use of any other resources or means than those indicated.

Osnabrück, August 27, 2009

(Daniel Berndt)