Towards a Corpus Annotated for Metonymies: the Case of Location Names

Katja Markert*, **Malvina Nissim***

*Division of Informatics
2 Buccleuch Place
EH8 9LW Edinburgh UK
{markert, malvi}@cogsci.ed.ac.uk

Abstract

At the moment, language resources do not contain the necessary information for large-scale metonymy processing. As a contribution, we here present a corpus annotated for metonymies. We describe a framework for annotating metonymies in domain-independent text that considers the regularity, productivity and underspecification of metonymic usage. We then present a fully worked out annotation scheme for location names and a gold standard corpus containing 2000 annotated location names. The annotation scheme is rigorously evaluated as to its reliability and compared to previous metonymy classification proposals. In particular, we show that it is not sufficient to rely on intuitions for reliable metonymy identification and that an annotation effort with trained annotators and explicit guidelines is necessary.

1. Introduction

Metonymy is a form of figurative speech, in which one expression is used to refer to the standard referent of a related one (Lakoff and Johnson, 1980). In Example (1), *"Vietnam"*, the name of a location, refers to an event (a war) that happened there.

(1) *"He was shocked by **Vietnam**."*

In Example (2),

(2) *"The **ham sandwich** is waiting for his check."*

"ham sandwich" refers to the customer who ordered the sandwich (Nunberg, 1978).

Metonymy (or regular polysemy) has generated high interest in linguistics and in natural language processing (NLP) as it is regular, productive and frequent (e.g. Markert and Hahn (2002) found a metonymy in 17% of all utterances in 27 German magazine texts). Metonymy resolution can also improve many language engineering tasks. Stallard (1993) cited a 27% performance improvement by incorporating metonymy resolution into a question answering system about a limited domain (commercial air flights), which had to understand metonymies such as *"Which **wide body jets** serve dinner?"*. Anaphora resolution, a crucial task in many NLP applications, often depends on metonymy recognition as well (Markert and Hahn, 2002; Harabagiu, 1998), as Example (3) from the Washington Post (Sunday, 28.10.2001) shows.

(3) *"**China** has agreed to let a United Nations investigator conduct an independent probe into .... But it was unclear whether **Beijing** would meet past UN demands for unrestricted access to .... "*

Here, the coreference chain can be established only if *"China"* and *"Beijing"* are recognised as metonymies for the government of China.

The main language resources, however, do not provide sufficient data about metonymy that could serve as a basis for large-scale testing of linguistic theories or NLP algorithms on naturally occurring texts.

Dictionaries necessarily include only conventional metonymic senses, whereas metonymies are open-ended, as Example (2) shows. But even conventional metonymic senses are often not included systematically. So "France", e.g., has one sense in WordNet (Fellbaum, 1998), the country, whereas "United States" has the additional metonymic sense "government of the US", which is clearly available for "France" as well. In addition, most dictionaries do not cover proper names, which can easily be used metonymically (Stern, 1931; Lakoff and Johnson, 1980).

Most **corpora** (e.g. the British National Corpus (BNC, `http://info.ox.ac.uk/bnc`)) do not contain any information about word senses. An example of a sense-annotated corpus is SEMCOR (Fellbaum, 1998), whose content words are tagged with their WordNet senses. Unfortunately, the shortcomings of dictionaries regarding metonymies are mirrored in the sense annotation — thus, "United States" is tagged with two distinct senses in SEMCOR, whereas "France" is always tagged with one sense only. In addition, SEMCOR's annotation might be unreliable: a replication experiment (Ng and Lee, 1996) showed only a 56% percentage agreement with the original annotation.

Most **example lists** in the linguistic literature (Stern, 1931; Lakoff and Johnson, 1980; Pustejovsky, 1995) contain only a small set of especially selected and/or constructed examples — thus, they do not provide an accurate picture of the range and distribution of phenomena in real-world texts. The authors also favour giving clear-cut examples, thus obscuring the fact that the literal/metonymic distinction might be hard to make reliably in practice.

This lack of language resources is the main cause of sparse evaluation of most NLP algorithms dealing with metonymy. Indeed, some of them are evaluated in comparison to constructed examples only (Utiyama et al., 2000; Fass, 1997; Hobbs et al., 1993; Pustejovsky, 1995), disregarding the range of phenomena in realistic settings. Others (Verspoor, 1997; Markert and Hahn, 2002; Harabagiu, 1998; Stallard, 1993) use naturally-occurring data that, however, seem to be analysed according to subjective intuitions of one individual only. These latter approaches seem to take for granted that the comparison data needed for their

algorithms (metonymies identified in natural language texts by humans) is easy to generate reliably, which presupposes that humans can easily agree on identification and interpretation of metonymies. Given experiences in sense annotation (Ng and Lee, 1996; Jorgensen, 1990), this seems unlikely as they show that disciplined efforts with several trained annotators are necessary to arrive at reliably annotated data (Kilgariff and Rosenzweig, 2000).

In this paper we address both the lack of language resources as well as the lack of annotation studies for metonymies. In particular,

- we present a general annotation framework for metonymies. This framework takes into account both technical desiderata (e.g., platfrom-independence) as well as linguistic properties of metonymies (e.g., regularity, productivity and underspecification);
- we describe a case study of location names to gather insights into abilities of humans to identify and interpret metonymies. We show that metonymy classifications as given in the linguistic literature are hard to reliably apply to real-world texts because the authors provide very limited information about the categories;
- we also show that this problem can be overcome by precise guidelines and trained annotators. We present a reliable and fully worked out annotation scheme for lcoation names, building on previous linguistic metonymy classifications but deviating when needed to improve reliability. The annotation scheme is evaluated with several experiments measuring annotation agreement of human judges;
- using this scheme we built an annotated gold standard corpus that includes 2000 literal and metonymic examples of location names, mirroring as far as possible the original distribution in a corpus of English texts.

The paper is organised as follows. In Section 2. we suggest a general framework for metonymy annotation. We then describe (Section 3.) how the data we use for our case study have been collected. In Section 4. we describe a first experiment carried out in order to test classifications from the literature. A fully worked out annotation scheme for location names that can serve as a blueprint for annotation schemes for other semantic classes is presented in Section 5.. Its reliability is rigorously evaluated in several reproducibility experiments described in Section 6. where we also present our gold standard corpus. We end the paper with discussions of related work (Section 7.) and our contributions (Section 8.).

2. Framework

We now present several principles for the construction of metonymy annotation schemes and annotated corpora.

The corpus should be annotated in a markup language that makes it reusable, platform-independent and easily searchable. We decided to use XML as it emerges as a standard in corpus markup for which searching and editing tools are available.

Principle 1 (Platform-independence) *Encode the corpus in XML.*

To make the corpus useful for many different applications we decided to include texts from as many different domains and genres as possible, hoping to cover a wide variety of metonymies. This is necessary as types and frequencies of metonymies can vary widely from genre to genre (in sports reports the use of a location name for a sports team (*"**Scotland** beat **Ireland**"*) is extremely frequent). Therefore we used the BNC, a 1 million word corpus that covers many domains and genres. All examples from now on are from the BNC, if not otherwise indicated by a *.

Principle 2 (Domain and genre) *Include as many different domain and genre types as possible.*

Traditionally, metonymy is seen as operating at the word level, extended to perhaps compounds as in Example (2) or multi-word names as "Republic of Germany" (see e.g. (Copestake and Briscoe, 1995)). Nunberg (1995), however, makes some convincing arguments for metonymy as a phrasal process, but to our knowledge no full account of the interaction of metonymy and phrasal semantics yet exists. Thus, we still attach any annotation to the head noun of the phrase. If the head noun is a compound or a multi-word name the annotation encompasses the whole compound/name.[1]

Principle 3 (Annotation extent) *The word level is the unit extent in annotation.*

Metonymic readings are very systematic (e.g. location names can be used productively for an associated event, as in Example (1)). Therefore, linguistic studies (Stern, 1931; Lakoff and Johnson, 1980; Fass, 1997) have postulated *metonymic patterns* (e.g. `place-for-event`) that operate on *semantic classes* (here, locations). Our annotation scheme will take advantage of such patterns in order to express regularities and ease annotation effort. Therefore, we developed general guidelines (specifying extent of annotation units, annotation procedure etc.) and specific guidelines for each semantic class covered (specifying metonymic patterns distinctive to this semantic class). The semantic classes we use are derived from both the metonymy literature and lexical databases like WordNet. Example classes are "location", "animal" and "plant".

Principle 4 (Regularities) *Use semantic classes and metonymic patterns for defining annotation categories.*

The intended referent of a metonymy is rarely as clear as in Example (2). In an example like *"**Hungary** took similar actions..."* it might be clear in context that Hungarian officials are involved but the decision-makers cannot be named exactly. Therefore we annotate just the *base class* of the noun — i.e., its original literal class, here "location" —

[1] The rest of the annotation scheme (e.g indicating the type of metonymy) is to a large degree independent of this decision, so that the annotation extent could be changed to phrasal annotation, if wished in the future.

and the metonymic pattern used, here e.g. `place-for-people`, which then implicitly gives the *intended class*, here "people". We annotate both base class and intended class as subsequent reference can refer to either as the example pair *"I bought a **Picasso**. He was a great painter."** and *"I bought a **Picasso**. It is a great painting."** shows.

Principle 5 (Underspecification) *Annotate both base class and metonymic pattern.*

Although the extensive use of metonymic patterns will greatly enhance the coverage of the annotation scheme, there must be at least one category for unconventional metonymies like Example (2).

Principle 6 (Coverage) *Cover conventional and unconventional metonymies.*

Figure 1 shows the basic XML-template for metonymy annotation and Figure 2 an example output for the class "location".

Grounding on Principle 4, this general framework has to be supplemented by specific guidelines/annotation schemes for semantic classes that undergo regular polysemy as these specify the annotation categories applicable. We conducted a case study concentrating on one semantic class, namely "location". Using only one class for the first human experiments minimises the annotation effort as the annotators do not have to "switch" between different annotation categories for different semantic classes.

3. Data Collection

In the long run, full text annotation (annotating all "locations" in a number of full texts) is most desirable in order to encounter the full range of possible phenomena (albeit dependent on the text genres) — but for this study we extracted text samples containing occurrences of location names that were taken from a previously compiled gazetteer (see below). The reason for this was twofold: firstly, it allows to collect more data in less time as otherwise one might have to annotate many texts without many words of the desired semantic class in them; secondly, and more importantly, we wanted to confine our experiments to testing the ability of humans to distinguish between metonymic and literal readings of location names without introducing the additional task of first identifying the location names in free text, which might lead to additional reliability problems.

For generating appropriate data for location names we extracted all country names from WordNet and the CIA factbook (`http://www.cia.gov/cia/publications/factbook/`). We then clustered different names referring to the same nation (e.g. greece|ellas|hellenic republic|elliniki dhimokratia for 'Greece'). This collection of names forms our sampling frame CountryList.

We built two sets of sample data, SetA for discussion/training and SetB for testing. In order to build SetA we randomly selected 10 country names from the CountryList and a total of 1000 occurrences (100 for each name) have been randomly sampled from the BNC. SetA used this stratified random sampling to test whether metonymic patterns identifiable for one country name can be carried over to any other country name. SetA has been subsequently split into SetA1 (300 samples), to be used for preliminary analysis of the linguistic classification systems in Experiment I, and SetA2 (700 samples) for training the annotators to use our final annotation scheme in Experiment II. In contrast, SetB contains 1000 occurrences of country names, randomly extracted from the BNC, allowing any country name in CountryList to occur. SetB has been used for testing our final annotation scheme in Experiment II.

We searched the BNC using Gsearch (Corley et al., 2001). All samples include three sentences of context.

4. Experiment I

We used SetA1 to test the reliability of the available linguistic classifications. In order to do so, we built a comprehensive record of metonymic patterns for locations by comparing and matching over 20 proposals from the literature and considering readings included in dictionaries. For example, we matched the Object-People pattern proposed by Copestake and Briscoe (1995) with the Place-Inhabitants one proposed by Stern (1931). The metonymic patterns `obj-for-name`, `obj-for-rep`, `place-for-org`, `place-for-off`, `place-for-pop`, `place-for-product` and `place-for-event` were included. We included all categories found in the literature (with occassionally different names) as well as a category `othermet` to handle unconventional metonymies. These categories were not structured any further and no preference rankings were given. We included two categories for literal readings, distinguishing the locative and the state/political reading of the name as it is e.g., done in WordNet. Last, three categories to remove noise introduced by our extraction method were provided. In Section 5. all the categories are described.

4.1. Method

*Annotators.*The annotators are two computational linguists and are the authors of this paper.

Guidelines and Training. The written guidelines consisted of very simple instructions and the list of categories given above. A range of examples from the literature and the BNC were included. Initially, the annotators independently annotated a set of 100 examples from SetA1. This was done without prior training. The results of this first exercise were then discussed and the remaining 200 examples of SetA1 were afterwards annotated.

Reliability Measures. Reproducibility of results (Krippendorff, 1980) has been measured using the kappa statistic (K), which corrects percentage agreement ($P(A)$) between annotators for expected chance agreement ($P(E)$) (Carletta, 1996).

4.2. Results

All results in this paper are rounded to the second decimal.

The annotation exercise performed without training produced a Kappa score of $K = .39$ $(N = 100; k = 2)$ where K stands for the Kappa coefficient, N for the number of examples annotated and k for the number of an-

<BASE-CLASS reading=*readingtype* metotype=*metopattern*> *annotated-noun* </BASE-CLASS> *continued-text* . . .

Figure 1: XML template for metonymy annotation

<LOCATION reading="metonymic" metotype="place-for-people"> Hungary </LOCATION> took similar actions . . .

Figure 2: XML output of annotation for the class "location"

notators. On Krippendorff's (1980) scale, agreement of $K \geq .80$ is considered as reliable, agreement between $.80$ and $.67$ as marginally reliable and lower agreement as unreliable. Thus, the results of our first exercise were highly unreliable. The second annotation exercise, performed on the remaining 200 examples of SetA1 after discussing the first, gave a better result with $K = .55$ $(N = 200; k = 2)$, but still unreliable. We discussed the problematic cases and found that many proposed distinctions from the literature were unclear even after training. In particular, some categories seemed too specific. In addition, the distinction between literal and metonymic readings was rarely as clear-cut as in the literature examples. Also, the syntactic structure of the sentences in our corpus was more complex than in the literature examples, often evoking literal and metonymic aspects of a name at the same time. This led to modify our annotation scheme towards a better articulated scheme, which is described in the following section along with the specific changes we made.

5. An Annotation Scheme for Location Names

Our extraction method can lead to collecting some undesired examples, i.e. *noise*. Thus, we postulate three categories, namely `unsure`, `nonapp`, and `homonym` that handle noise and have to be considered before any further annotation. These were also included in the initial scheme.

Rarely, the limited amount of context we extracted prevented the annotator from understanding the sample. Whenever this is the case, the category `unsure` must be applied and no further analysis performed.

Following MUC-7 (Chinchor, 1997), we regard proper names as atomic. Sometimes an extracted name N (as *"UK"* in Example (4)) is part of a complex proper name denoting a different entity (*"Shell UK"* in Example (4)). The name N is to be annotated as `nonapp` and no further annotation will be performed.

(4) *"Shell **UK**"*

The country names in CountryList can also be used as names for other semantic classes. In these cases, a different base class and a category `homonym` are assigned (see Example (5)).

(5) *"Rear Admiral **Poland**"*

If the sample is understood and the extracted name atomic and of the desired base class, the annotation can proceed to identify *literal*, *metonymic*, and *mixed* readings.

The `literal` reading for location names comprises a locative (see Example (6)) and a political entity interpretation (see Example (7)).

(6) *"coral coast of **Papua New Guinea**"*

(7) *"**Britain**'s current account deficit"*

The locative and the political sense (often distinguished in dictionaries as well as in our initial scheme) frequently proved hard to distinguish in our data, as Example (8) illustrates. Here, the unions are both legally affiliated to the state Britain as well as locally situated in the country. Therefore we merged these two readings to one literal reading.

(8) *"**Britain**'s unions"*

For metonymic readings, we distinguish between *general* patterns (valid for all physical objects) and *location-specific* ones. General patterns are:

- `obj-for-rep`: the name refers to a representation of the standard referent (photo, painting, etc.), as in Example (9).

  (9) *"This is **Malta**"** (pointing to a map)

- `obj-for-name`: the name is used as a mere signifier. In Example (10), *"Guyana"* would receive a literal interpretation, whereas *"British Guiana"* is a mere reference to a previous name.

  (10) *"Guyana (formerly **British Guiana**) gained independence"*

Location-specific patterns are:

- `place-for-people`: a place stands for any persons/organisations associated with it. Often, the explicit referent is underspecified, as in Example (11), where the reference could be to the government, an organisation or the whole population.

  (11) *"The G-24 group expressed readiness to provide **Albania** with food aid"*

  It is therefore important to assign the right pattern (`place-for-people`) at a higher level, and a more specific pattern (*subtype*), if identifiable, at a lower level. Such a hierarchical approach has the great advantage of 'punishing' disagreement only at a later stage and allowing fall-back options for automatic systems. This leads to Principle 7 to be integrated into our general framework.

Principle 7 (Hierarchical structure) *Organise the categories hierarchically.*

Note that in the literature and our original scheme the patterns are not hierarchically organised, thus not showing relations between them.

We therefore introduce four optional subtypes for the `place-for-people` pattern.

`Cap-Gov` (only for capitals of countries/states) identifies a capital standing for the government of the whole country (*"**Rome** decided...*"*).

`Off` identifies the official administration (e.g. the government or the army).

Examples are given in (12) and (13).

(12) *"EC denunciations of **Israel**'s handling of the intifada"*

(13) *"**America** did once try to ban alcohol"*

`Org` identifies organisations (or a set of organisations) associated with the location (including sport teams, companies, etc.); a list of possible organisations has been extracted from WordNet. In Example (14), *"San Marino"* identifies the national football team. In Example (15), *"France"* refers to college(s) located in France.

(14) *"a 29th-minute own goal from **San Marino** defender Claudio Canti"*

(15) *"Mr Peter Shuker, the principal, said the college now had links with **France**"*

`Pop` identifies the whole or majority of the population, as in the religious context of Example (16).

(16) *"The notion that the incarnation was to fulfil the promise to **Israel** and to reconcile the world with God"*

- `place-for-event`: a location name stands for something that there happened or the current situation of the location. This category is usually illustrated with very clear-cut examples in the literature, but it proved difficult to distinguish from literal readings in practice. (This was also due to its extreme rarity, which did not help in singling out relevant clues). For example, the occurrence of *"Bosnia"* in Example (17) clearly refers to the war there, but the occurrence of *"Sweden"* in Example (18) is less clear-cut. Indeed, the reference (in this particular context) was to a sports event in Sweden, but the literal reading is still true and the metonymic `place-for-event` one can be obtained by inference. In such cases, we opt for `literal`, introducing a preference between readings that was not present in our initial scheme.

  (17) *"you think about some of the crises that are going on in the world from **Bosnia** and so on"* (`place-for-event`)

  (18) *"he didn't play in **Sweden**"* (`literal`)

- `place-for-product`: a place stands for a product there manufactured (e.g. *"Bordeaux*" can refer to the wine there produced).

The category `othermet` covers unconventional metonymies (see Principle 6). Since they are open-ended and context-dependent, no specific category indicating the intended class can be introduced. In Example (19), *"New Jersey"* metonymically refers to the local typical tunes. The category `othermet` is only used if none of the other categories fits.

(19) *"The thing about the record is the influences of the music. The bottom end is very New York/**New Jersey** and the top is very melodic"*

In addition to literal and metonymic readings, we found examples where two predicates are involved, triggering a different reading each, thus yielding a *mixed* reading. This occurs very often with *coordinations* and *appositions*.

(20) *"they arrived in **Nigeria**, hitherto a leading critic of ..."*

In Example (20), both a literal (triggered by "arriving in") and a `place-for-people` (with subtype `Off`) reading (triggered by "leading critic") are invoked. We therefore introduced the category `mixed` to deal with these cases (not treated as a category in the literature).

Figure 3 summarises the scheme. Summarising our changes to the categories proposed in the literature, we

- merged categories that could not be distinguished reliably;
- organised categories hierarchically when they were related and underspecified readings were frequent;
- proposed preference rankings between categories that can be easily confused;
- introduced the category `mixed` to handle cases where two readings are evoked by different predicates in the same sentence.

6. Experiment II

We describe here the experiment carried out to test the reliability of our annotation scheme for location names.

6.1. Method

Annotators. The annotators are the authors of this paper.

Guidelines and Training. The written annotation guidelines consist of general guidelines (containing annotation extent and instructions for `nonapp`, `unsure`, `homonym`, `obj-for-rep` and `obj-for-name`) and guidelines for the metonymic patterns specific to the semantic class "location" (see also Table 3). Identification of readings is driven by *replacement tests* described in the guidelines (e.g. if an occurrence of "Vietnam" can be replaced by "the war in Vietnam", we annotate it as `place-for-event`). The

<table>
<tr><th>Understanding</th><th>App</th><th>Base-type</th><th>Reading</th><th>Pattern</th><th>Subtype</th></tr>
<tr><td colspan="6">unsure</td></tr>
<tr><td rowspan="9">yes</td><td colspan="5">no</td></tr>
<tr><td rowspan="8">yes</td><td colspan="4">homonym</td></tr>
<tr><td rowspan="8">location</td><td colspan="3">literal</td></tr>
<tr><td rowspan="6">metonymic</td><td colspan="2">obj-for-rep</td></tr>
<tr><td colspan="2">obj-for-name</td></tr>
<tr><td colspan="2">place-for-event</td></tr>
<tr><td rowspan="2">place-for-people</td><td rowspan="2">CapGov
Off
Org
Pop</td></tr>
<tr></tr>
<tr><td colspan="2">place-for-product</td></tr>
<tr><td colspan="2">othermet</td></tr>
<tr><td colspan="3">mixed</td></tr>
</table>

Figure 3: Annotation Scheme — Text-understanding, Applicability, Readings, Metonymic Patterns

guidelines also contain examples for each category and instructions for ambiguous and underspecified cases. The annotators have been trained by independently annotating SetA2, containing 700 occurrences of 10 country names (see Section 3.). The annotation was performed using the MATE annotation tool (Isard et al., 2000) that was customised for metonymy annotation.

Test. SetB — containing 1000 occurrences of country names, allowing every country name to occur — was used for testing annotation. Again, no discussion was allowed during annotation.

Reliability Measures. We evaluated the reproducibility of results again by using the kappa statistic (K).

6.2. Results

A summary of the results is given in Table 1.

Reproducibility disregarding subtypes. We measured reproducibility of the distinctions between the categories `unsure`, `nonapp`, `homonym`, `literal`, `obj-for-rep`, `obj-for-name`, `place-for-event`, `place-for-people`, `place-for-product`, `othermet` and `mixed`, thereby disregarding the level of subtypes. This set of supertype categories will be called `Supertypes` from now onwards. Reproducibility for the testset SetB was measured at $K = .88$ ($N = 1000; k = 2$). Thus, the results can be considered as reliable annotation.

Reproducibility including subtypes. We measured reproducibility of the extended set of categories, consisting of the union of `Supertypes` and the subtypes `CapGov`, `Off`, `Org` and `Pop`. A lower (still reliable) $K = .81$ ($N = 1000; k = 2$) matched the assumption that subtypes are harder to distinguish than the supertypes level.

Reproducibility excluding noise. The annotation categories as considered until now also include those identifying noisy data. These categories were experienced as easier to apply than the literal/metonymy distinctions. To test whether reproducibility was still sufficient for these harder distinctions we measured reproducibility on a subset of the test set SetB containing only the examples that no annotator marked as `unsure`, `nonapp` or `homonym`. Measuring reproducibility using the remaining categories in `Supertypes` yielded $K = .87$ ($N = 931, k = 2$). Extending this category set with subtypes yielded $K = .78$ ($N = 931, k = 2$). Thus the sense distinction on the supertype level is still reliable; incorporating the subtypes induces a substantial drop to marginal reliability (although this number can still be considered high in the field of sense annotation). This shows the virtues of a hierarchical scheme with cut-off points.

Single category reliability. The experience of the annotators seemed to indicate that some categories (such as subtypes) are harder to identify than others (such as `nonapp`). We used Krippendorff's (1980) single category reliability to discover which categories the human judges found difficult to identify. For a single category, agreement is measured by collapsing all categories but the one of interest into one meta-category and then calculating kappa as usual. As for some categories data was sparse and therefore results could be misleading, we only measured single category reliability for categories that were used at least 10 times by the annotators (e.g. 6 times by one annotator and 7 by the other). These categories are `nonapp`, `literal`, `place-for-people`, `mixed`, `othermet` and the subtypes `Off`, `Org` and `Pop`. As expected, `nonapp` was easiest to identify (see Table 1 for all results). Also reliable are the annotations of the most frequent readings `literal` and `place-for-people`. The annotation of `mixed` readings was only marginally reliable. A plausible explanation is that their identification involves the correct recognition of at least two categories. Reproducibility for `othermet` was also marginally reliable, showing that the identification of unconventional metonymies is harder than the ones with predefined intended classes. Regarding subtypes, the reliability for `Org` was high, also due to the large number of easily identifiable sports teams examples in our data. The reproducibility of `Off` was marginally reliable, whereas `Pop` was the only category that was unreliably annotated. The reason for this is that it is often hard to decide whether the population, the government or an organisation is involved (see again Example 11) — it even depends to some degree of the picture of political influence the annotator has.

Influence of training. To quantify the influence of training we measured reproducibility on the training set SetA2, using `Supertypes` only. The result was $K = .80$ ($N = 700, k = 2$), which was substantially lower than the corresponding result (.88) on the testset SetB. This shows the importance of training for achieving high agreement. On the other hand, it shows that training on some specific words of a semantic class might be sufficient for annotating all words from this class (remember that the training set contained only 10 different country names).

Gold standard corpus. After the annotation we developed a gold standard for the testset, discussing the cases we had not agreed on in the annotation exercise. Of the 1000 examples, 61 (6.1%) were excluded as noise. 737 (73.7%) examples are literal, 161 (16.1%) `place-for-people` metonymies, 3 (.3%) event metonymies, 9 (.9%) were annotated as `othermet` and 15 (1.5%) as `mixed`. Neither `obj-for-rep` nor `obj-for-name` nor `place-for-product` were found in the testset SetB. Only 14 (1.4%) cases could not be agreed on even after discussion. Figure 4 shows the distribution of readings in SetB. Regarding the subtypes of the 161 `place-for-people` metonymies, 79 could be identified as `Off`, 41 as `Org` and only 4 as `Pop` examples. In 37 cases, no subtype could be identified.[2]

7. Related Work

There are not many corpus studies for metonymies. Automatic algorithms for both *recognition* and *interpretation* of metonymies that are evaluated on corpora are presented in (Stallard, 1993; Harabagiu, 1998; Markert and Hahn, 2002). However, none of them seems to use principled annotation schemes for identifying metonymies, thus limiting the evaluation to subjective comparison. Lapata (2001) evaluates an interpretation algorithm by letting human judges rank its results. However, she only treats *logical metonymy* of the easily identifiable pattern "adjective-noun" (e.g. "fast car", "fast secretary") and does not handle metonymy recognition that we concentrate on. Verspoor (1997) also restricts herself to interpretation, but she does not compare her intuitions to other annotators' judgments.

The only other metonymy annotation scheme we know of is being developed within the ACE project (`http://www.itl.nist.gov/iad/894.01/tests/ace/`). Besides locations, it covers organisations, facilities, and persons, but only very few metonymic patterns are used. For locations, only equivalents to our subtypes of `place-for-people` (with no hierarchical structure) are included. No categories for *mixed* and unconventional readings exist. Agreement data has not been published yet.

8. Conclusions and Future Directions

We have presented a corpus of 2000 occurrences of location names annotated with regard to literal and metonymic usage. We have also described a general annotation framework for metonymies which is applicable to other semantic classes and which takes linguistic properties of metonymies into account.

Our annotation scheme for location names covers the metonymic patterns presented in the literature and enhances them by introducing explicit guidelines and preference rankings that allow reliable annotation, by introducing a category `mixed` for cases where different readings are invoked simultaneously and by structuring categories hierarchically. The latter improvement ensures *progressive sense refinement* (Resnik and Yarowsky, 2000), allowing automatic systems fall-back options.

We have also tested the ability of humans to reliably identify metonymies and shown that intuitive judgements following informal linguistic definitions/classifications are not sufficient and that training as well as explicit guidelines are necessary for reliable metonymy annotation. We have described several annotation experiments, showing very good reproducibility results for our annotation scheme. Our choice of hierarchical organisation is supported by the results obtained for the category `place-for-people` and its subtypes. We have also shown the positive effect that training has on reliability.

In the future, we intend to proceed to full text annotation and to extend our guidelines to cover other semantic classes, e.g., "person" and "organisation".

Acknowledgements. Katja Markert is funded by an Emmy Noether Fellowship of the Deutsche Forschungsgemeinschaft (DFG) and Malvina Nissim by ESRC Project R000239444. We thank Amy Isard for her customisation of the MATE annotation workbench.

9. References

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Nancy Chinchor. 1997. MUC-7 Named Entity Task definition. In *Proc. of the 7^{th} Conference on Message Understanding; 1997*, Washington, DC.

Ann Copestake and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

Steffan Corley, Martin Corley, Frank Keller, Matthew Crocker, and Shari Trewin. 2001. Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. *Computers and the Humanities*, 35(2):81–94.

Dan Fass. 1997. *Processing metaphor and Metonymy*. Ablex, Stanford, CA.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Sanda Harabagiu. 1998. Deriving metonymic coercions from WordNet. In *Workshop of the Usage of WordNet in Natural Language Processing Systems, COLING-ACL '98*, pages 142–148, Montreal, Canada.

Jerry R. Hobbs, Mark E. Stickel, Douglas E. Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

Amy Isard, David McKelvie, Andreas Mengel, and Morten Baun Moller. 2000. The MATE Workbench – an annotation tool. In Maria Gavrilidou, Geroge Carayan-

[2] We also agreed on a gold standard for the 1000 occurrences in SetA so that the whole gold standard corpus contains 2000 examples. Due to lack of space we do not discuss the SetA distribution here.

Set	Use	Categories	N	P(A)	P(E)	K
SetA2	train	All supertypes	700	.94	.69	.80
SetB	test	All supertypes	1000	.95	.58	.88
SetB	test	All supertypes and subtypes	1000	.92	.56	.81
SetB (no noise)	test	Supertypes (no noise)	931	.95	.66	.87
SetB (no noise)	test	Supertypes + Subtypes (no noise)	931	.92	.64	.78
SetB	test	single:nonapp	1000	.99	.91	.96
SetB	test	single:literal	1000	.95	.62	.88
SetB	test	single:people	1000	.97	.73	.90
SetB	test	single:mixed	1000	.99	.97	.76
SetB	test	single:othermet	1000	.99	.98	.73
SetB	test	single:off	1000	.97	.86	.76
SetB	test	single:org	1000	.98	.92	.81
SetB	test	single:pop	1000	.99	.986	.57

Table 1: Reliability results for Experiment II

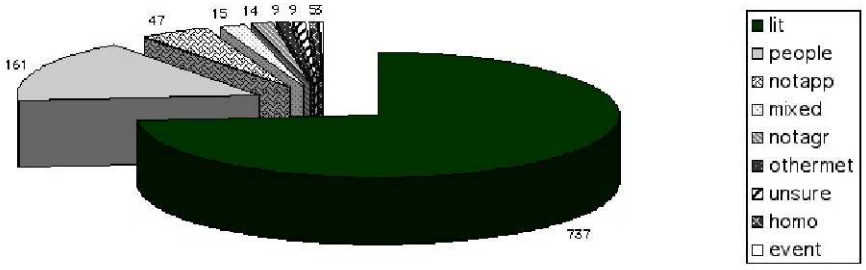

Figure 4: Distribution of Readings in SetB

nis, Stella Markantonatou, Stelios Piperdis, and Gregory Stainhaouer, editors, *Proc. of the 2^{nd} International Conference on Language Resources and Evaluation; Athens, Greece, 2000*, pages 1565–1570, Athens, Greece.

Julia Jorgensen. 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19(3):167–190.

Adam Kilgariff and Joseph Rosenzweig. 2000. English senseval: Report and results. In *Proc. of the 2^{nd} International Conference on Language Resources and Evaluation; Athens, Greece, 2000*, Athens, Greece.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. Chicago University Press, Chicago, Ill.

Maria Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proc. of the 2nd Meeting of the North American Chapter of the ACL*, Pittsburgh, PA.

Katja Markert and Udo Hahn. 2002. Understanding metonymies in discourse. *Artificial Intelligence*, 135(1/2):145–198, February.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proc. of the 34^{th} Annual Meeting of the Association for Computational Linguistics; Santa Cruz, Cal., 23–28 June 1996*, pages 40–47, Santa Cruz, Ca.

Geoffrey Nunberg. 1978. *The Pragmatics of Reference*. Ph.D. thesis, City University of New York, New York.

Geoffrey Nunberg. 1995. Transfers of meaning. *Journal of Semantics*, 12:109–132.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, Mass.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

David Stallard. 1993. Two kinds of metonymy. In *Proc. of the 31^{st} Annual Meeting of the Association for Computational Linguistics; Columbus, Ohio, 22-26 June 1993*, pages 87–94, Columbus, Ohio.

Gustav Stern. 1931. *Meaning and Change of Meaning*. Göteborg: Wettergren & Kerbers Förlag.

Masao Utiyama, Masaki Murata, and Hitoshi Isahara. 2000. A statistical approach to the processing of metonymy. In *Proc. of the 18^{th} International Conference on Computational Linguistics; Saarbruecken, Germany, 1–4 August 2000*, pages 885–891, Saarbrücken, Germany.

Cornelia Verspoor. 1997. Conventionality-governed logical metonymy. In H. Bunt, L. Kievit, R. Muskens, and N. Verlinden, editors, *Proc. of the 2nd International Workshop on Computational Semantics*, pages 300–312, Tilburg, The Netherlands.