Running head: Corpus-based Metonymy Analysis

Corpus-based Metonymy Analysis

Katja Markert and Malvina Nissim

Division of Informatics

University of Edinburgh

Abstract

In this paper we make the case for corpus-based metonymy analysis and show that many interesting linguistic and statistical questions can only be answered by working with real texts. To facilitate such studies, we present a method for annotating metonymies in domain and genre-independent text. We advocate an annotation scheme that builds on regularities in metonymic usage, that takes underspecification in metonymic reference into account and that is organised hierarchically. We combine previous metonymy classification proposals with insights from a corpus study to present a fully worked-out annotation scheme for location names, illustrating the above principles. We present several experiments measuring annotation agreement and show that the annotation scheme is reliable and has wide coverage. We also provide a gold standard for annotations of this kind consisting of 2000 annotated occurrences of country names in the British National Corpus. We use the resulting corpus to study metonymy distributions and the factors that influence the choice of literal vs. metonymic readings in real texts.

Corpus-based Metonymy Analysis

Metonymy is a form of figurative speech, in which one expression is used to refer to the standard referent of a related one (Lakoff & Johnson, 1980). So, in

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

"Vietnam", the name of a location, refers to an event (a war) that happened there. In

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

"ham sandwich" refers to the customer who ordered the sandwich (Lakoff & Johnson, 1980; Stallard, 1993).

As metonymy has generated considerable interest in linguistics (Stern, 1931; Lakoff & Johnson, 1980; Nunberg, 1995; Pustejovsky, 1995; Panther & Radden, 1999), by now, several characteristics of metonymies have been brought to light and interesting claims have been made:

1. Metonymic readings are very systematic: for example, location names can be productively used to refer to an associated event (see Example (1) and similar examples like Woodstock). Linguistic studies (Stern, 1931; Lakoff & Johnson, 1980; Fass, 1997) have therefore postulated conventionalised metonymic patterns (e.g. `place-for-event`) that operate on semantic classes (here, locations).

2. Unconventional metonymies (see Example (2)) can be created on the fly. Their interpretation is context-dependent.

3. Metonymy seems to be quite frequent.

However, the insights above stem from studies that are based mainly on linguistic intuition, instead of corpus studies, and are often biased to make a particular point of

interest (for example, stressing metonymic patterns over unconventional metonymies or vice versa). Therefore, such studies are only illustrated by small sets of especially selected and/or constructed examples, cover only a limited range of what might be encountered in real-world texts and do not necessarily provide an accurate picture of the actual distribution of phenomena. Specifically, they leave the following questions still unanswered:

1. What is the actual distribution of literal readings, conventional metonymic patterns and unconventional metonymies in real-world texts?

2. Which factors (e.g., text type or word-specific behaviour) influence the distribution of metonymies?

3. How valid are the pattern lists proposed in the literature, regarding their coverage and granularity when applied to real text?

The well-formedness of the above questions relies on the assumption that literal and metonymic readings as well as different metonymic patterns can be reliably distinguished. The clear-cut examples normally given in the literature hide the fact that such distinctions might be hard to make in practice. Indeed, it is unclear whether the proposed pattern lists can serve as an annotation scheme for metonymy markup by humans. In most work on metonymy, metonymy identification is based on (undocumented) human intuitions (Stallard, 1993; Pustejovsky, 1995; Fass, 1997; Markert & Hahn, 2002, e.g.), which have proved unreliable in other areas of sense annotation (Jorgensen, 1990; Ng & Lee, 1996).

In order to answer these questions a large amount of natural language data analysed for metonymy is needed, but is unfortunately not yet available. This paper describes such a corpus analysis of metonymies in English texts, centering on the following points:

1. Development of a reliable annotation scheme for literal vs. metonymic usage. The annotation scheme builds on the linguistic insights mentioned so far and uses metonymic patterns defined on semantic classes for its annotation categories. We extensively tested

reproducibility, coverage and granularity implications of this scheme.

2. Use of the developed annotation scheme to build a 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.

3. Exploration of the distribution of metonymies in the corpus.

4. Determination of some of the factors influencing the choice of literal and metonymic usage, making use of the developed gold standard corpus.

In the next section we present our fully worked out annotation scheme for location names that can serve as a blueprint for annotation schemes for other semantic classes. Its reliability is rigorously evaluated in the evaluation section. We then present our gold standard corpus and discuss its distribution in the light of the above questions. As possible factors influencing metonymy distribution we discuss the information coded in a semantic class vs. the information coded in a particular lemma, the influence of textual domain and the influence of the "level" of writing. We end the paper with discussions of related work and our contributions.

An Annotation Scheme for Location Names

In this paper we concentrate on an annotation scheme for the semantic class of "locations", which illustrates all properties of our general annotation framework (Markert & Nissim, 2002).

Data Collection

We extracted all country names from WordNet (Fellbaum, 1998) and the CIA factbook[1]. This collection of names constitutes our sampling frame CountryList.

We sampled occurrences of these country names from the British National Corpus (BNC)[2], a 100 million word corpus of contemporary British English that covers many

domains and genres. All examples from now on are from the BNC, if not otherwise indicated by a *.

We built two sets of sample data: SetA and SetB. 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 extracted from the BNC. SetA used this stratified random sampling in order to allow comparisons of reading distributions for different names. SetA has been subsequently split into SetA1 (300 samples), to be used for preliminary analysis of metonymic patterns, and SetA2 (700 samples) for training the annotators in using our final annotation scheme. In contrast, SetB contains 1000 occurrences of country names, randomly extracted from the BNC, allowing any country name in the CountryList to occur. SetB has been used for testing our final annotation scheme.

We searched the BNC using Gsearch (Corley, Corley, Keller, Crocker, & Trewin, 2001). All samples in both sets include two sentences of context before and one after the extracted name.

Annotation Scheme

We built a comprehensive record of metonymic patterns for locations by comparing and matching over 20 proposals from the literature (Lakoff & Johnson, 1980; Stern, 1931, e.g.) and considering metonymic 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). We tested this pattern collection for annotation reliability using SetA1 (Markert & Nissim, 2002) and refined and enhanced it, leading to the final annotation scheme described below.

Our extraction method can collect some undesired examples, i.e. noise. We therefore postulate three categories handling noise that have to be considered before any

further annotation: `unsure`, `nonapp`, and `homonym`.

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 "US" in Example (3)) is part of a complex proper name denoting a different entity ("US Open" in (3)). The name N is to be annotated as `nonapp` and no further annotation will be performed.

(3) "**US** Open"

The country names in CountryList can also be used as names for other semantic classes. In these cases, a different semantic class and a category `homonym` are assigned to the name N (see Example (4)).

(4) "Rear Admiral **Poland**"

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

The `literal` reading for location names comprises a locative and a political entity interpretation, as in (5) and (6), respectively. The division between the locative and the political sense which is often proposed in dictionaries is very hard to distinguish in our data, as the example "**Britain**'s unions" illustrates.

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

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

For metonymic readings, we postulate general patterns that are 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 (7).

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

– `obj-for-name`: the name is used as a mere signifier. In Example (8), "Guyana" would receive a literal interpretation whereas "British Guiana" is a mere reference to a previous name of the country.

(8) " 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 left underspecified, as in Example (9), where the reference could be to the government, an organisation or the whole population.

(9) "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 to annotation categories has the great advantage of 'punishing' disagreement only at a later stage, allowing fall-back options in annotation. Note that in the literature up to now the patterns are not hierarchically organised, thus not showing the relations between them.

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

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

`Off` identifies the official administration (including the government, the army, etc.). An example is given in (10).

(10) "**America** did . . . try to ban alcohol"

`Org` identifies an organisation (or a set of organisations) associated with the location; a list of possible organisations has been extracted from WordNet. In Example (11), "San Marino" identifies the national football team. In Example (12), "France" refers to college(s) located in France.

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

(12) "Mr Peter Shuker [. . .] said the college now had links with **France**"

`Pop` identifies the whole/majority of the population, as in the religious context that characterises Example (13).

(13) "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 happened there. This category is illustrated with very clear-cut examples in the literature, but proved difficult to distinguish from literal readings in practice. For example, whereas the occurrence of "Bosnia" in Example (14) clearly refers to the war in Bosnia, the occurrence of "Sweden" in (15) is less clear-cut. Indeed, the context points to a sport competition in Sweden, but the literal reading is still true and the metonymic `place-for-event` can be obtained by inference. In such ambiguous cases, we opt for a literal reading.

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

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

– `place-for-product`: the place stands for a product manufactured there; an example is the use of "Bordeaux" to refer to the wine produced there.

The category `othermet` covers unconventional metonymies that can refer to an open-ended set of semantic classes. In Example (16), "New Jersey" metonymically refers to typical tunes from New Jersey. The category `othermet` is used only if none of the other categories fits.

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

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 especially often with coordinations and appositions.

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

In Example (17), both the literal (triggered by "arriving in") and a `place-for-people` reading (triggered by "leading critic") are invoked. In order to deal with these cases (not treated as a category in the literature) we introduced the category `mixed`.

Figure 1 summarises the whole scheme.

Insert Figure 1 about here

Evaluation

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

Method

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

Guidelines and Training. The written guidelines for annotation consist of general guidelines (containing instructions for `nonapp`, `unsure`, `homonym`, `obj-for-rep` and `obj-for-name`) and guidelines for the location-specific metonymic patterns. 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", a `place-for-event` reading is to be annotated). The guidelines also contain examples for each category and instructions for ambiguous cases[3].

The annotators have been trained by independently annotating SetA2. The annotation was performed using the MATE annotation tool (Isard, McKelvie, Mengel, & Moller, 2000).

Test. SetB was used for testing annotation. Again, no discussion was allowed during annotation.

Reliability Measures. We evaluated the reproducibility of results (Krippendorff, 1980) by using the kappa statistic (K), which measures agreement among a set of annotators making category judgements, correcting for expected chance agreement (Carletta, 1996):

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of times the annotators agree, and $P(E)$ is the proportion of times they are expected to agree by chance.

Results

All results are rounded to the second decimal.

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` (see Figure 1). This set of supertype categories will be called `Supertypes` from now onwards. Reproducibility for the testset SetB was measured at $K = .88$ ($N = 1000; k = 2$ where K stands for the Kappa coefficient, N for the number of examples annotated and k for the number of annotators). 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 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`. We expected Kappa to drop as subtypes are hard to distinguish. Agreement, though still reliable, was measured at $K = .81$ ($N = 1000; k = 2$).

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. Measuring reproducibility using the no-noise categories in `Supertypes` only 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 where cut-off points are possible.

Single category reliability. We used Krippendorff's (1980) single category reliability to discover which categories the human judges found easy or hard 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 ($K = .96, N = 1000, k = 2$). Also reliable are the annotations of the most frequent readings `literal` ($K = .88, N = 1000, k = 2$) and `place-for-people` ($K = .90, N = 1000, k = 2$). The annotation of `mixed` readings was only marginally reliable ($K = .76, N = 1000, k = 2$) , probably because their identification involves the recognition of at least two categories. Reproducibility for `othermet` was also marginally reliable ($K = .73, N = 1000, k = 2$), showing that the identification of unconventional metonymies is harder. Regarding subtypes, the reliability for `Org` ($K = .81, N = 1000, k = 2$) was high due to the large number of easily identifiable sports teams examples in our data. The reproducibility of `Off` was marginally reliable ($K = .76, N = 1000, k = 2$), whereas `Pop` was the only category that was unreliably annotated ($K = .57, N = 1000, k = 2$). The reason for this drop in reliability is that it is often hard to decide whether the whole population, the government or some organisation is involved.

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. This shows the importance of training for achieving high agreement.

The Distribution of Metonymies and its Determining Factors

After the annotation process we agreed on a gold standard for SetB as well as for SetA, discussing the cases on which we did not agree during annotation. Thus, the

combined gold standard corpora contain 2000 annotated examples.[4] These gold standard corpora now allow for the study of the distribution of metonymies and of the factors that might influence the choice of metonymic or literal usage.

Frequency of metonymies. Of the 1000 examples of country names in SetB, 61 (6.1%) were excluded as noise. 14 (1.4%) examples could not be agreed on even after discussion. 737 (73.7%) examples are literal, 161 (16.1%) are `place-for-people` metonymies, 3 (.3%) `place-for-event` metonymies, 9 (.9%) were annotated as `othermet` and 15 (1.5%) as mixed readings. Neither `obj-for-rep` nor `obj-for-name` nor `place-for-product` were found in SetB.

The distribution is therefore highly skewed towards literal and `place-for-people` readings. The most surprising result is the low number of `place-for-event` metonymies, which are often emphasized in pattern lists in the literature. This might be due to the fact that we regarded country names only, whereas e.g. cultural events are frequently associated with cities. The scheme achieved a good coverage with only 0.9% of examples being annotated as unconventional metonymies. This shows that even though the kinds of metonymic reference are limitless in theory (see Example (2)), in practice some high-frequency metonymic references constitute the overwhelming majority of readings. Regarding subtypes, of the 161 `place-for-people` metonymies, 79 could be identified as `Off`, 41 as `Org` (of which 30 were sports team metonymies, and a further 9 metonymic readings for business organisations) and only 4 as `Pop` examples. In 37 cases, no subtype could be identified, showing that underspecified readings are quite frequent.

Word-based vs. class-based frequencies. The frequency studies on SetB give us information about the reading distribution on all country names, thereby taking a class-based approach. Although the metonymic patterns operate on whole classes, one of the factors that influence the frequency of metonymies, however, might be the particular name and its associations. Therefore we would expect, for example, a higher frequency of

`place-for-event` metonymies for countries where famous events happened (e.g., Vietnam) than for countries on average.

We tested this hypothesis on SetA (containing 10 sets, each of which in turn contains 100 examples of occurrences of one particular country name). We compared the proportion of nonliteral vs. literal readings for each pair of country names, thereby performing 45 comparisons, and tested the significance of the difference in proportion using a standard z-test with significance level 0.05. 17 of the 45 comparisons resulted in significant differences. In general, countries that are economically or politically more important to a UK reader (e.g., Japan) have a higher proportion of metonymic readings than less important countries (e.g., Guyana) , due to a higher number of `place-for-people` metonymies. Applications (like automatic metonymy recognition) would do well to take the differences in distribution for individual names into account.

Text type. It is highly likely that different text types contain different kinds as well as a different number of metonymies. We explored this hypothesis in our corpus by studying how both differences in domain and differences in the level of writing change the percentage of nonliteral readings in texts. All percentage differences were again tested for significance using a z-test with significance level 0.05.

As a first hypothesis, we explored whether there is a difference in metonymy frequency for imaginative vs. informative texts. We used the BNC classification of all written texts into imaginative and informative texts and measured the percentage of metonymic readings in both categories. Contrary to our intuition there was no significant difference between the two text types. However, this result might be caused by one or more of the following factors:

1. The BNC domain classification was carried out subjectively by the BNC annotators and reliability figures were not issued and/or computed.

2. The number of samples from imaginative texts is very low (38). More data from

this text type might be necessary.

3. The influence of the text domain might be dependent of the type of metonymies studied. The metonymies mainly studied in this paper are highly conventional and often appear in informative texts.

As a second hypothesis we tested whether the level of writing influences metonymy frequency. We again used the BNC classification which divides all written texts into three "Levels", an assessment of the text's technicality or difficulty, with Level 1 being the easiest and Level 3 the most difficult. Examples of Level 1 texts are texts from the newspaper "Daily Mirror" or from a hair magazine called HAIR FLAIR 1992. Examples of Level 3 texts are books like "Russia in the age of reaction and reform 1801 -1881" or "The political economy of soil erosion in developing countries". Again, the assessment of difficulty was a subjective judgement of BNC annotators.

We compared the frequency of metonymies in all Levels pairwise. The differences between Level 2 and Level 3 were not significant. In contrast, the differences between Level 1 and Level 2 as well as between Level 1 and Level 3 were significant. Easier texts contained significantly fewer metonymies than harder texts: only about 10% of all occurrences of country names in Level 1 texts were metonymic in contrast to about 25% in Level2/3 texts. This is interesting as the metonymies studied are very conventional and yet do not frequently occur in very simple texts.

Related Work

The most prominent language ressources existing for metonymies are dictionaries and corpora. **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, the country, whereas United States has the additional metonymic sense

"government of the US". Our annotation scheme, on the contrary, works systematically for a whole semantic class and therefore has no such inconsistencies. In addition, most dictionaries do not cover proper names.

Like the BNC, most existing **corpora** 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. Therefore, SEMCOR does not serve as a good searchable source of metonymy occurrences. In addition, SEMCOR's annotation might be unreliable (Ng & Lee, 1996), whereas our corpus has been evaluated stringently for its annotation reliability.

There are not many corpus studies for metonymies (see the Introduction). Some mainly domain-specific distribution and frequency studies are mentioned in (Stallard, 1993; Verspoor, 1997; Harabagiu, 1998; Markert & Hahn, 2002). However, none of them seems to use principled annotation schemes for identifying metonymies, thus limiting the evaluation to subjective 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/`). Their annotation scheme also uses a class- and pattern-based approach. In addition to locations, organisations, facilities, and persons are covered, but only a very limited number of metonymic patterns is used. For locations, only equivalents to our subtypes of `place-for-people` (without a hierarchical structure) are included. No categories for `mixed` readings and unconventional metonymies exist. Agreement data has not been published yet.

Conclusions

We have made the case for corpus-based metonymy analysis to test linguistic claims about metonymy.

As a basis for corpus studies, we presented a fully worked out annotation scheme for location names that is compatible with previous more informal metonymy classifications. It enhances them by introducing a category `mixed` for dealing with cases where different readings are invoked simultaneously, and by structuring categories hierarchically. The latter improvement ensures progressive sense refinement (Resnik & Yarowsky, 2000), the basic semantic class (e.g., location) distinguishing approximately at the level of homonyms, the metonymic patterns at the level of regular polysemy and metonymic subtypes further specifying intended referent types. This allows fall-back options in corpus studies.

We have also described several annotation experiments, showing very good results for reproducibility. Our choice for hierarchical organisation has been validated by the results obtained for the category `place-for-people` and its subtypes.

We created a reliably annotated corpus that includes a wide range of genres and styles. We have shown how such a corpus can be used to evaluate coverage and granularity of metonymy classifications as well as for studying frequency distributions and the factors that influence the choice of literal or metonymic usage. We think that many linguistic theories can and should be tested on such real world data and that our corpus is a first step in providing a corpus that allows for larger scale testing.

In the future we intend to enlarge our annotated corpus, and to extend our annotation scheme to other semantic classes, so that we can proceed to full text annotation.

References

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

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

Copestake, A., & Briscoe, T. (1995). Semi-productive polysemy and sense extension. Journal of Semantics, 12, 15-67.

Corley, S., Corley, M., Keller, F., Crocker, M., & Trewin, S. (2001). Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. Computers and the Humanities, 35(2), 81-94.

Fass, D. (1997). Processing metaphor and metonymy. Stanford, CA: Ablex.

Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. Cambridge, Mass.: MIT Press.

Harabagiu, S. (1998). Deriving metonymic coercions from WordNet. In Workshop on the Usage of WordNet in Natural Language Processing Systems, COLING-ACL '98 (p. 142-148). Montreal, Canada.

Isard, A., McKelvie, D., Mengel, A., & Moller, M. B. (2000). The MATE Workbench – an annotation tool. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperdis, & G. Stainhaouer (Eds.), Proc. of the 2^{nd} International Conference on Language Resources and Evaluation; Athens, Greece, 2000 (p. 1565-1570).

Jorgensen, J. (1990). The psychological reality of word senses. Journal of Psycholinguistic Research, 19(3), 167-190.

Krippendorff, K. (1980). Content analysis: An introduction to its methodology. Sage Publications.

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. Chicago, Ill: Chicago University Press.

Markert, K., & Hahn, U. (2002). Understanding metonymies in discourse. Artificial Intelligence, 135(1/2), 145-198.

Markert, K., & Nissim, M. (2002). Towards a corpus annotated for metonymies: the case of location names. In Proc. of the 3^{rd} International Conference on Language Resources and Evaluation; Las Palmas, Canary Islands, 2002.

Ng, H. T., & Lee, H. B. (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 (p. 40-47).

Nunberg, G. (1995). Transfers of meaning. Journal of Semantics, 12, 109-132.

Panther, K.-U., & Radden, G. (Eds.). (1999). Metonymy in thought and language. Benjamins.

Pustejovsky, J. (1995). The generative lexicon. MIT Press, Cambridge, Mass.

Resnik, P., & Yarowsky, D. (2000). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. Natural Language Engineering, 5(2), 113-133.

Stallard, D. (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 (p. 87-94).

Stern, G. (1931). Meaning and change of meaning. Göteborg: Wettergren & Kerbers Förlag.

Verspoor, C. (1997). Conventionality-governed logical metonymy. In H. Bunt, L. Kievit, R. Muskens, & N. Verlinden (Eds.), Proc. of the 2nd International Workshop on Computational Semantics (p. 300-312). Tilburg, The Netherlands.

Author Note

Katja Markert and Malvina Nissim, Division of Informatics, University of Edinburgh.

This study is supported by grants from the ESRC, UK, and the DFG, Germany.

Correspondence should be addressed to Katja Markert, Division of Informatics, University of Edinburgh, 2 Buccleuch Place, EH8 9LW Edinburgh, UK, markert@inf.ed.ac.uk.

Footnotes

[1] http://www.cia.gov/cia/publications/factbook/

[2] http://info.ox.ac.uk/bnc

[3] A public version of the guidelines is available at
http://www.cogsci.ed.ac.uk/~malvi/mascara/publications.html

[4] The corpora are freely available in XML format — please contact the authors for a copy.

Figure Captions

Figure 1. Annotation Scheme—Text Understanding, Applicability, Readings, Metonymic Patterns

Comprehension	App	Base-type	Reading	Pattern	Subtype
unsure					
	no				
		homonym			
			literal		
				obj-for-rep	
				obj-for-name	
		location		place-for-event	
	yes		metonymic		CapGov
				place-for-people	Off
					Org
					Pop
				place-for-product	
				othermet	
			mixed		