Detecting Article Errors Based on the Mass Count Distinction

Ryo Nagata[1], Takahiro Wakana[2], Fumito Masui[2], Atsuo Kawai[2], and Naoki Isu[2]

[1] Hyogo University of Teacher Education, 942-1 Shimokume, Yashiro, 673-1494 Japan
rnagata@info.hyogo-u.ac.jp
[2] Mie University, 1577, Kurimamachiya, Tsu, 514-8507, Japan
{wakana, masui, kawai, isu}@ai.info.mie-u.ac.jp

Abstract. This paper proposes a method for detecting errors concerning article usage and singular/plural usage based on the mass count distinction. Although the mass count distinction is particularly important in detecting these errors, it has been pointed out that it is hard to make heuristic rules for distinguishing mass and count nouns. To solve the problem, first, instances of mass and count nouns are automatically collected from a corpus exploiting surface information in the proposed method. Then, words surrounding the mass (count) instances are weighted based on their frequencies. Finally, the weighted words are used for distinguishing mass and count nouns. After distinguishing mass and count nouns, the above errors can be detected by some heuristic rules. Experiments show that the proposed method distinguishes mass and count nouns in the writing of Japanese learners of English with an accuracy of 93% and that 65% of article errors are detected with a precision of 70%.

1 Introduction

Although several researchers [1,2,3] have shown that heuristic rules are effective to detecting grammatical errors in the English writing of second language learners, it has been pointed out that it is hard to write heuristic rules for detecting article errors [1]. To be precise, it is hard to write heuristic rules for distinguishing mass and count nouns which are particularly important in detecting article errors. The major reason for this is that whether a noun is a mass noun or a count noun greatly depends on its meaning or its surrounding context (Refer to Pelletier and Schubert [4] for detailed discussion on the mass count distinction).

Article errors are very common among Japanese learners of English [1,5]. This is perhaps because the Japanese language does not have an article system similar to that of English. Thus, it is favorable for error detecting systems aiming at Japanese learners of English to be capable of detecting article errors. In other words, such systems need to somehow distinguish mass and count nouns in the writing of Japanese learners of English.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 815–826, 2005.

In view of this background, we propose a method for automatically distinguishing mass and count nouns in context to complement the conventional heuristic rules for detecting grammatical errors. In this method, mass and count nouns are distinguished by words surrounding the target noun. Words surrounding the target noun are collected from a corpus and weighted based on their occurrences. The weighted words are used for distinguishing mass and count nouns in detecting article errors.

Given the mass count distinction, errors concerning singular/plural usage, which are also common in the writing of Japanese learners of English, can be detected as well as article errors. For example, given that the noun *information* is a mass noun, *informations* can be detected as an error. Considering this, we include errors concerning singular/plural usage in the target errors of this paper. Hereafter, to keep the notation simple, the target errors[1] will be referred to as article errors.

The next section describes related work on distinguishing mass and count nouns. Section 3 proposes the method for automatically distinguishing mass and count nouns. Section 4 describes heuristic rules for detecting article errors based on the mass count distinction given by the proposed method. Section 5 discusses results of experiments conducted to evaluate the proposed method.

2 Related Work

Several researchers have proposed methods for distinguishing mass and count nouns in the past. Allan [6] has presented an approach to distinguishing nouns that are used only as either mass or count based on countability environments. This distinction is called countability preferences. Baldwin and Bond [7,8] have proposed several methods for learning the countability preferences from corpora[2]. Bond and Vatikitis-Bateson [9] have shown that nouns' countability can be predicted using an ontology[3]. O'Hara et al. [10] have proposed a method for classifying mass and count nouns based on semantic information (Cyc ontological types [11]).

Unfortunately, it is difficult to apply the above methods to complement the conventional heuristic rules for detecting grammatical errors. The methods [6,7,8,9] are not enough for the purpose, because the majority of nouns can be used as both mass and count depending on the surrounding context [12]. The methods [9,10] cannot be readily applicable to the purpose because they work only when semantic information on nouns is given. It would be difficult to extract semantic information from nouns in the writing of learners of English.

[1] The details of the target errors are shown in Sect. 4.

[2] They define four way countability preferences: fully countable, uncountable, bipartite, and plural only.

[3] They define five way countability preferences: fully countable, strongly countable, weakly countable, uncountable, and plural only.

3 Distinguishing Mass and Count Nouns

In the proposed method, decision lists [13] are used to distinguish mass and count nouns. Generally, decision lists are learned from a set of manually tagged training data. In the proposed method, however, training data can be automatically generated from a raw corpus.

Section 3.1 describes how to generate training data. Section 3.2 describes how to learn decision lists from the training data. Section 3.3 explains the method for distinguishing mass and count nouns using the decision lists.

3.1 Generating Training Data

To generate training data, first, instances of the target noun that head their noun phrase (NP) are collected from a corpus with their surrounding words. This can be simply done by an existing chunker or parser.

Then, the collected instances are tagged with mass or count by tagging rules. For example, the underlined *chicken*:

Example 1. ... are a lot of *chicken*s in the roost ...

is tagged as

Example 2. ... are a lot of *chicken*s/*count* in the roost ...

because it is in plural form.

We have made tagging rules based on linguistic knowledge [6,14,12]. Figure 1 and Table 1 represent the tagging rules. Figure 1 shows the framework of the tagging rules. Each node in Fig. 1 represents a question applied to the instance in question. For example, the root node reads "Is the instance in question plural?". Each leaf represents a result of the classification. For example, if the answer is 'yes' at the root node, the instance in question is tagged with count. Otherwise, the question at the lower node is applied and so on. The tagging rules do not classify instances as mass or count in some cases. These unclassified instances are tagged with the symbol '?'. Unfortunately, they cannot readily be included in training data. For simplicity of implementation, they are excluded from training data.

Table 1. Words used in the tagging rules

(a)	(b)	(c)
the indefinite article	much	*the definite article*
another	less	*demonstrative adjectives*
one	enough	*possessive adjectives*
each	all	*interrogative adjectives*
—	sufficient	*quantifiers*
—	—	*'s genitives*

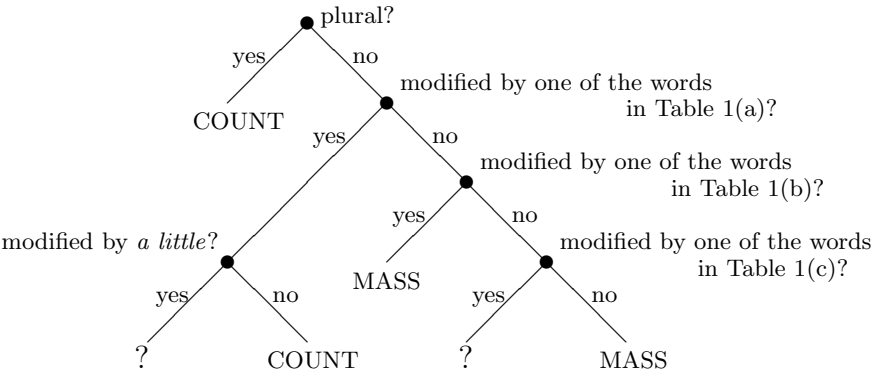

Fig. 1. Framework of the tagging rules

Note that the tagging rules can be used only for distinguishing mass and count nouns in texts containing no errors. They cannot be used in the writing of Japanese learners of English that may contain errors including article errors; they are based on article and the distinction between singular and plural.

Finally, the tagged instances are stored in a file with their surrounding words. Each line in the file consists of one of the tagged instances and its surrounding words as shown in *Example 2*. The file is used as training data for learning a decision list.

3.2 Learning Decision Lists

A decision list consists of a set of rules that are learned from training data. Each rule matches with the template as follows:

$$\text{If } a \textit{ condition is true}, \text{ then } a \textit{ decision}. \tag{1}$$

To define the template in the proposed method, let us have a look at the following two examples:

Example 3. I read the *paper*.

Example 4. The *paper* is made of hemp pulp.

The underlined *paper*s in both sentences cannot simply be classified as mass or count by the tagging rules presented in Sect. 3.1 because both are singular and modified by the definite article. Nevertheless, we can tell that the former is a count noun and that the latter is a mass noun from the contexts. This suggests that the mass count distinction is often determined by words surrounding the target noun. In *Example 3*, we can tell that the *paper* refers to something that can be read from *read*, and therefore it is a count noun. Likewise, in *Example 4*, the *paper* refers to a certain substance from *made* and *pulp*, and therefore it is a mass noun.

Taking this observation into account, we define the template based on words surrounding the target noun. To formalize the template, we will use a random variable MC that takes either *mass* or *count* to denote that the target noun is a mass noun or a count noun, respectively. We will also use w and C to denote a word and a certain context around the target noun, respectively. We define three types of C: np, $-k$, and $+k$ that denote the contexts consisting of the noun phrase that the target noun heads, k words to the left of the noun phrase, and k words to its right, respectively. Then the template is formalized by

> If a word w appears in the context C of the target noun,
> then the target noun is distinguished as MC.

Hereafter, to keep the notation simple, the template is abbreviated to

$$w_C \rightarrow MC. \tag{2}$$

Now rules that match with the template can be learned from the training data generated in Sect. 3.1. All we need to do is to collect words in C from the training data. Here, the words in Table 1 are excluded. Also, function words such as pronouns and auxiliary verbs, cardinal and quasi-cardinal numerals, and the target noun are excluded. All words are reduced to their morphological stem and converted entirely to lower case when collected. For example, the following tagged instance:

Example 5. She ate a piece of fried *chicken*/mass for dinner.

would give a set of rules that match with the template:

Example 6.
$piece_{-3} \rightarrow mass, \quad fry_{np} \rightarrow mass, \quad dinner_{+3} \rightarrow mass$

for the target noun *chicken* being *mass* when $k = 3$.

In addition to the above rules, a default rule is defined. It is based on the target noun itself and used when no other confident rules[4] are found in the decision list for the target noun. It is defined by

$$t \rightarrow MC_{\text{major}} \tag{3}$$

where t and MC_{major} denote the target noun and the major case of MC in the training data, respectively. Equation (3) reads "If the target noun appears, then it is distinguished as the major case".

The log-likelihood ratio [15] decides in which order rules in a decision list are applied to the target noun in novel context. It is defined by

$$log \frac{p(MC|w_C)}{p(\overline{MC}|w_C)} \tag{4}$$

[4] Confidence is given by the log-likelihood ratio, which will be defined by (4).

where $\overline{MC}$ is the exclusive event of MC and $p(MC|w_C)$ is the probability that the target noun is used as MC when w appears in the context C. For the default rule, the log-likelihood ratio is defined by

$$log\frac{p(MC_{\text{major}}|t)}{p(\overline{MC_{\text{major}}}|t)} \tag{5}$$

It is important to exercise some care in estimating $p(MC|w_C)$. In principle, we could simply count the number of times that w appears in the context C of the target noun used as MC in the training data. However, this estimate can be unreliable, when w does not appear often in the context. To solve this problem, using a smoothing parameter α [16], $p(MC|w_C)$ is estimated by

$$p(MC|w_C) = \frac{f(w_C, MC) + \alpha}{f(w_C) + m\alpha} \tag{6}$$

where $f(w_C)$ and $f(w_C, MC)$ are occurrences of w appearing in C and those in C of the target noun used as MC, respectively. The constant m is the number of possible classes, that is, $m = 2$ (*mass* or *count*) in our case, and introduced to satisfy $p(MC|w_C) + p(\overline{MC}|w_C) = 1$. In this paper, α is set to 0.5. Likewise, $p(MC_{\text{major}}|t)$ is estimated by

$$p(MC_{\text{major}}|t) = \frac{f(t, MC_{\text{major}}) + \alpha}{f(t) + m\alpha} \tag{7}$$

Rules in a decision list are sorted in descending order by (4) and (5). They are tested on the target noun in novel context in this order. Rules sorted below the default rule are discarded because they are never used as we will see in Sect. 3.3.

Table 2 shows part of a decision list for the target noun *chicken* that was learned from a subset of the BNC (British National Corpus) [17]. Note that the rules are divided into two columns for the purpose of illustration in Table 2; in practice, they are merged into one just as shown in Table 3.

On one hand, we associate the words in the left half with food or cooking. On the other hand, we associate those in the right half with animals. From this observation, we can say that *chicken* is a count noun in the sense of an animal but a mass noun when referring to food or cooking, which agrees with the knowledge presented in previous work [18].

Table 2. Rules in a decision list (target noun: *chicken*, $k = 3$)

Mass		Count	
w_C	Log-likelihood ratio	w_C	Log-likelihood ratio
$piece_{-3}$	1.49	$count_{-3}$	1.49
$fish_{-3}$	1.28	$peck_{+3}$	1.32
$dish_{-3}$	1.23	pig_{np}	1.23
$skin_{+3}$	1.23	run_{-3}	1.23
$serve_{+3}$	1.18	egg_{np}	1.18

Table 3. An example of a decision list (target noun: *chicken*, $k = 3$)

Rules	Log-likelihood ratio
$\underline{piece_{-3} \rightarrow mass}$, $count_{-3} \rightarrow count$	1.49
$peck_{+3} \rightarrow count$	1.32
$fish_{-3} \rightarrow mass$	1.28
$dish_{-3} \rightarrow mass$, $pig_{np} \rightarrow count$, $\cdots$	1.23
:	:

3.3 Distinguishing the Target Noun in Novel Context

To distinguish the target noun in novel context, each rule in the decision list is tested on it in the sorted order until the first applicable one is found. It is distinguished by the first applicable one. If two or more applicable rules (e.g., "$piece_{-3} \rightarrow mass$" and "$count_{-3} \rightarrow count$" in Table 3) are found, it is distinguished by the major decisions of the two or more applicable rules. For example, suppose there are three applicable rules and two of them are for mass nouns (one of them is for count nouns). In this case, the target noun is distinguished as mass. Ties are broken by rules sorted below the ties. If ties include the default rule, it is distinguished by the default rule.

The following is an example of distinguishing the target noun *chicken*. Suppose that the decision list shown in Table 3 and the following sentence are given:

Example 7. I ate a piece of *chicken* with salad.

It turns out that the first rule "$piece_3 \rightarrow mass$" in Table 3 is applicable to the instance. Thus, it is distinguished as a mass noun.

It should be noted that rules sorted below the default rule are never used because the default rule is always applicable to the target noun. This is the reason why rules sorted below the default rule are discarded as mentioned in Sect. 3.2.

4 Heuristic Rules for Detecting Article Errors

So far, a method for distinguishing mass and count nouns has been described. This section describes heuristic rules for detecting article errors based on the mass count distinction given by the method.

Article errors are detected by the following three steps. Rules in each step are examined on each target noun in the target text.

In the first step, any mass noun in plural form is detected as an article error. If an article error is detected in the first step, the rest of the steps are not applied.

In the second step, article errors are detected by the rules described in Table 4. The symbol "$\star$" in Table 4 denotes that the combination of the corresponding row and column is erroneous. For example, the third row denotes that

Table 4. Detection rules used in the second step

Pattern	Count		Mass	
	Singular	Plural	Singular	Plural
{another, each, one}	—	⋆	⋆	⋆
{a lot of, all, enough, lots of, sufficient}	⋆	—	—	⋆
{much}	⋆	⋆	—	⋆
{kind of, sort of, that, this}	—	⋆	—	⋆
{few, many, these,those}	⋆	—	⋆	⋆
{countless, numerous, several, various}	⋆	—	⋆	⋆
cardinal number except one	⋆	—	⋆	⋆
{any, some, no, *'s genitives*}	—	—	—	⋆
{*interrogative adjectives*, *possessive adjectives*}	—	—	—	⋆

Table 5. Detection rules used in the third step

	Singular			Plural		
	a	the	ϕ	a	the	ϕ
Mass	⋆	–	–	⋆	⋆	⋆
Count	–	–	⋆	⋆	–	–

plural count nouns, singular mass nouns, and plural mass nouns that are modified by *another*, *each*, or *one* are erroneous. The symbol "—" denotes that no error can be detected by the table. If one of the rules in Table 4 is applied to the target noun, the third step is not applied.

In the third step, article errors are detected by the rules described in Table 5. The symbols "a", "the", and "ϕ" in Table 5 denote the indefinite article, the definite article, and no article, respectively. The symbols "⋆" and "—" are the same as in Table 4. For example, "⋆" in the third row and second column denotes that the singular mass nouns modified by the indefinite article is erroneous.

In addition to the three steps, article errors are detected by exceptional rules. The indefinite article that modifies other than the head noun is judged to be erroneous (e.g., *an expensive). Likewise, the definite article that modifies other than the head noun and adjectives is judged to be erroneous (e.g., *the them).

5 Experiments

5.1 Experimental Conditions

A subset of essays[5] written by Japanese learners of English were used as the target texts in the experiments. The subset contained 30 essays (1747 words). A native speaker of English who was a professional rewriter of English recognized 62 article errors in the subset.

[5] http://www.lb.u-tokai.ac.jp/lcorpus/index-j.html

The British National Corpus (BNC) [17] was used to learn decision lists. Spoken data were excluded from the corpus. Also, sentences the OAK system[6], which was used to extract NPs from the corpus, failed to analyze were excluded. After these operations, the size of the corpus approximately amounted to 80 million words (the size of the original BNC is approximately 100 million words). Hereafter, unless otherwise specified, the corpus will be referred to as the BNC.

Performance of the proposed method was evaluated by accuracy, recall, and precision. Accuracy is defined by

$$\frac{\text{No. of mass and count nouns distinguished correctly}}{\text{No. of distinguished target nouns}}. \tag{8}$$

Namely, accuracy measures how accurately the proposed method distinguishes mass and count nouns. Recall is defined by

$$\frac{\text{No. of article errors detected correctly}}{\text{No. of article errors in the target essays}}. \tag{9}$$

Recall measures how well the proposed method detects all the article errors in the target essays. Precision is defined by

$$\frac{\text{No. of article errors detected correctly}}{\text{No. of detected article errors}}. \tag{10}$$

Precision measures how well the proposed method detects only the article errors in the target essays.

5.2 Experimental Procedures

First, decision lists for each target noun in the target essays were learned from the BNC. To extract noun phrases and their head nouns, the OAK system was used[7]. An optimal value for k (window size of context) was estimated as follows. For 23 nouns[8] shown in [12] as examples of nouns used as both mass and count nouns, accuracy was calculated using the BNC and ten-fold cross validation. As a result of setting $k = 3, 10, 50$, it turned out that $k = 3$ maximized the average accuracy. Following this result, $k = 3$ was selected in the experiments.

Second, the target nouns were distinguished whether they were mass or count by the proposed method, and then article errors were detected by the mass

[6] OAK System Homepage: http://nlp.cs.nyu.edu/oak/

[7] We evaluated how accurately training data can be generated by the tagging rules using the OAK system. It turned out that the accuracy was 0.997 against 2903 instances of 23 nouns shown in [12] which were randomly selected from the BNC; 1694 of those were tagged with mass or count by the tagging rules and 1689 were tagged correctly. The five errors were due to the OAK system.

[8] In [12], 25 nouns are shown. Of those, two nouns (*hate* and *spelling*) were excluded because they only appeared 12.1 and 15.6 times on average in the ten-fold cross validation, respectively.

count distinction and the heuristic rules described in Sect. 4. As a preprocessing, spelling errors in the target essays were corrected using a spell checker.

Finally, the results of the detection were compared to those done by the native-speaker of English. From the comparison, accuracy, recall, and precision were calculated.

Comparison of performance of the proposed method to that of other methods is difficult because there is no generally accepted test set or performance baseline [19]. Given this limitation, we compared performance of the proposed method to that of Grammarian[9], a commercial grammar checker. We also compared it to that of a method that used only the default rules in the decision lists. We tested them on the same target essays to measure their performances.

5.3 Experimental Results and Discussion

In the experiments, the proposed method distinguished mass and count nouns in the target essays with accuracy of 0.93. This means that the proposed method is effective to distinguishing mass and count nouns in the writing of Japanese learners of English. From this result, we can say that the proposed method can complement the conventional heuristic rules for detecting grammatical errors.

Because of the high accuracy of the proposed method, it detected more than half of the article errors in the target essays (Table 6). Of the undetected article errors (22 out of 62), only four were due to the misclassification of mass and count nouns by the proposed method. The rest were article errors that were not detected even if the mass count distinction was given. For example, extra definite articles such as "I like *the gardening." cannot be detected even if whether the noun "gardening" is a mass noun or a count noun is given. Therefore, it is necessary to exploit other sources of information than the mass count distinction to detect these kinds of article error. For instance, exploiting the relation between sentences could be used to detect these kinds of article error.

The proposed method outperformed the method using only the default rules in both recall and precision. This means that words surrounding the target nouns are good indicators of the mass count distinction. For example, the proposed method correctly distinguished the target noun *place* in the phrase *beautiful place* as a count noun by "$beautiful_{np} \rightarrow count$" and detected an article error from it whereas the method using only the default rules did not.

Table 6. Experimental results

Method	Recall	Precision
Proposed	0.65	0.70
Default only	0.60	0.69
Grammarian	0.13	1.00

[9] Grammarian Pro X ver. 1.5: http://www.mercury-soft.com/

In precision, the proposed method was outperformed by Grammarian; since Grammarian is a commercial grammar checker, it seems to be precision-oriented. The proposed method made 17 false-positives. Of the 17 false-positives, 13 were due to the misclassification of mass and count nouns by the proposed method. Especially, the proposed method often made false-positives in idiomatic phrases (e.g., by plane). This result implies that some methods for handling idiomatic phrases may improve the performance. Four were due to the chunker used to analyze the target essays. Since the chunker is designed for analyzing texts that contain no errors, it is possible that a chunker designed for analyzing texts written by Japanese learners of English reduces this kind of false-positive.

6 Conclusions

This paper has proposed a method for distinguishing mass and count nouns to complement the conventional heuristic rules for detecting grammatical errors. The experiments have shown that the proposed method distinguishes mass and count nouns with a high accuracy (0.93) and that the recall and precision are 0.65 and 0.70, respectively. From the results, it follows that the proposed method can complement the conventional heuristic rules for detecting grammatical errors in the writing of Japanese learners of English.

The experiments have also shown that approximately 35% of article errors in the target essays are not detected by the mass count distinction. For future work, we will study methods for detecting the undetected article errors.

Acknowledgments

The authors would like to thank Sekine Satoshi who has developed the OAK System. The authors also would like to thank three anonymous reviewers for their advice on this paper.

References

1. Kawai, A., Sugihara, K., Sugie, N.: ASPEC-I: An error detection system for English composition. IPSJ Journal (in Japanese) **25** (1984) 1072–1079
2. McCoy, K., Pennington, C., Suri, L.: English error correction: A syntactic user model based on principled "mal-rule" scoring. In: Proc. 5th International Conference on User Modeling. (1996) 69–66
3. Schneider, D., McCoy, K.: Recognizing syntactic errors in the writing of second language learners. In: Proc. 17th International Conference on Computational Linguistics. (1998) 1198–1204
4. Pelletier, F., Schubert, L.: Two theories for computing the logical form of mass expressions. In: Proc.10th International Conference on Computational Linguistics. (1984) 108–111

5. Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., Isahara, H.: Automatic error detection in the Japanese learners' English spoken data. In: Proc. 41st Annual Meeting of the Association for Computational Linguistics. (2003) 145–148
6. Allan, K.: Nouns and countability. J. Linguistic Society of America **56** (1980) 541–567
7. Baldwin, T., Bond, F.: A plethora of methods for learning English countability. In: Proc. 2003 Conference on Empirical Methods in Natural Language Processing. (2003) 73–80
8. Baldwin, T., Bond, F.: Learning the countability of English nouns from corpus data. In: Proc. 41st Annual Meeting of the Association for Computational Linguistics. (2003) 463–470
9. Bond, F., Vatikiotis-Bateson, C.: Using an ontology to determine English countability. In: Proc. 19th International Conference on Computational Linguistics. (2002) 99–105
10. O'Hara, T., Salay, N., Witbrock, M., Schneider, D., Aldag, B., Bertolo, S., Panton, K., Lehmann, F., Curtis, J., Smith, M., Baxter, D., Wagner, P.: Inducing criteria for mass noun lexical mappings using the Cyc KB, and its extension to WordNet. In: Proc. 5th International Workshop on Computational Semantics. (2003) 425–441
11. Lenat, D.: CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM **38** (1995) 33–38
12. Huddleston, R., Pullum, G.: The Cambridge Grammar of the English Language. Cambridge University Press, Cambridge (2002)
13. Rivest, R.: Learning decision lists. Machine Learning **2** (1987) 229–246
14. Gillon, B.: The lexical semantics of English count and mass nouns. In: Proc. Special Interest Group on the Lexicon of the Association for Computational Linguistics. (1996) 51–61
15. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proc. 33rd Annual Meeting of the Association for Computational Linguistics. (1995) 189–196
16. Yarowsky, D.: Homograph Disambiguation in Speech Synthesis. Springer-Verlag (1996)
17. Burnard, L.: Users Reference Guide for the British National Corpus. version 1.0. Oxford University Computing Services, Oxford (1995)
18. Ostler, N., Atkins, B.: Predictable meaning shift: Some linguistic properties of lexical implication rules. In: Proc. of 1st SIGLEX Workshop on Lexical Semantics and Knowledge Representation. (1991) 87–100
19. Chodorow, M., Leacock, C.: An unsupervised method for detecting grammatical errors. In: Proc. 1st Meeting of the North America Chapter of the Association for Computational Linguistics. (2000) 140–147