Generating a Category Set of Words Using a Hierarchical Part-of-Speech System and Tagged Corpus

Takeyuki KOJIMA and **Yoshiyuki KOTANI**
Dept. of Computer, Information and Communication Science,
Tokyo University of Agric. and Tech.
2-24-16 Nakacho, Koganei, Tokyo, Japan
{khojima,kotani}@fairy.ei.tuat.ac.jp

Abstract

In this paper, we propose a method of generating a proper categorization of morphemes by giving a hierarchical part-of-speech system and a corpus tagged using this part-of-speech system. Our method use hierarchical information in the part-of-speech system and statistical information in the corpus to generate a category set. The statistical information is based on the context of occurrence of categories. First, we specify the format of given information. Then, we describe an algorithm to generate a proper categorization. Finally, we present the results of our experiments in applying this method. We obtaind a moderately proper categorization and found several candidates for improvement.

1 Introduction

In natural language processing, it is important to categorize words or morphemes properly. A proper categorization depends on, among other things, the kind of processing task, the domain of target corpus, and the size of the corpus. When categorization is too general, we can not use characters of individual categories. Because the characters of categories hide each other. On the contrary, when categorization is too specific, we can not use the characters of categories also. Because a low frequency of a category decreases the reliability of characters of the category.

Past researches have proposed categorizations and tagsets for differnt purposes: morphological analysis, syntactic analysis or information extraction, etc. While some of the categorizations are made by hand using linguistic knowledge, others are created from annotated corpora automatically or semiautomatically. Several researches have forcused on methods of modifing the existing categorization in order to improve the accuracy of their task with respect to their purpose (Brants, 1995). Criteria of categorizations in these researches are the accuracies of their tasks. Other researches have proposed an criteria of categorizations of words based on linguistic quality and not processing quality (Déjean, 2000).

We propose here a method to decide a proper categorization of morphemes, giving a hierarchical part-of-speech system and a corpus tagged using this part-of-speech system. In other words, our method forcuses on reducing an existing category set using hierarchical information of part-of-speech system and statistical information of the corpus. We recursively subdivide the categories using topdown approach with subdivision score. The subdivision score, which indicates how significant it is to subivide a category, is based on difference between the context of the category and that of its parent category.

We explain structure of given information, which consist of a part-of-speech system and a tagged corpus, and a generating category set in Section 2. Section 3 explains the method of generating a category set. Section 4 shows experiments that are performed in order to test the generating algorithm. In the next section, we disscuss the result of the experiments.

2 Given Information and Generated Category Set

2.1 Hierarchical Part-of-Speech System

The part-of-speech system we use in this research is a tree structure specified by a triplet $U = \langle V_{\mathrm{M}}, V_{\mathrm{P}}, P \rangle$, where V_{M} is a set of the leaf part-of-speech categories of this tree structure. V_{P} is a set of the part-of-speech categories occupying the intermediate (or non-leaf) nodes including the root node ρ. In other words, each element of V_{M} is one of the most specific categories in the part-of-speech system. For instance, ρ-N-NO, ρ-N-PN-NT and ρ-SYM are elements of V_{M}, while ρ-N, ρ-N-PN and ρ are elements of V_{P}[1]. A hyphen stands for a parent-child relationship in these category names.

The parent-child relationship between the nodes of a tree structure is defined by the parent function $P : V_{\mathrm{M}} \cup V_{\mathrm{P}} \rightarrow V_{\mathrm{P}}$. A category $P(x)$ is more general than a category x. For instance, $P(\rho\text{-N-PN}) = \rho\text{-N}$ and $P(\rho\text{-SYM}) = \rho$. The parent of the root node is undefined.

Given a parent function P, the children function C is defined as follows :

$$C(x) = \{ y \,|\, P(y) = x \}.$$

$C(x)$ is a set of all children of a category x.

2.2 Tagged Corpus

The tagged corpus we use in this research is tagged with the part-of-speech system described in Section 2.1. The corpus is tagged with the most specific categories. In other words, each word in the original corpus is replaced with the leaf category of the part-of-speech system.

2.3 Category Set

Our objective is to generate a category set G from a part-of-speech system U and a tagged corpus, which is a subset of the category set of U; that is, $G \subset V_{\mathrm{M}} \cup V_{\mathrm{P}}$. Figure 1 shows a hierarchical part-of-speech system U and a category set G. In this figure, a circle stands for category and a line stands for the parent-child relationship between categories.

Since an input tagged corpus consists of leaf categories, all leaf categories must be replaced by corresponding categories that are elements of category set G. We put requirements on G, in order to ensure that all leaf categories have its ancestor category in G.

We recursively define a function δ as follows:

$$\delta^{U,G}(x) \equiv \begin{cases} x, & x \in G \\ \delta^{U,G}(P(x)), & \text{otherwise} \end{cases}$$

This equation means that $\delta^{U,G}(x)$ is an ancestor of x and it is an element of G. We require that the generated category set G must satisfy the following two requirements: (1) $\delta^{U,G}(x)$ is defined for all $x \in V_{\mathrm{M}}$, (2) $P(x) \notin G$ for all $x \in G$.

2.4 Environment of a Category

We quantify the context of occurrence in a corpus Π with a category set G by the terms $p_{\mathrm{L}}^{\Pi,U,G}(x;z)$ and $p_{\mathrm{R}}^{\Pi,U,G}(x;z)$ as follows:

$$p_{\mathrm{L}}^{\Pi,U,G}(x;z) \equiv \frac{f_G^{\Pi,U}(x \cdot z)}{f_G^{\Pi,U}(z)}$$

$$p_{\mathrm{R}}^{\Pi,U,G}(x;z) \equiv \frac{f_G^{\Pi,U}(z \cdot x)}{f_G^{\Pi,U}(z)}.$$

[1] N, NO, PN, RG, NT and SYM stand for noun, number, propernoun, region, nation and symbol, respectively.

In these equations, $f_G^{\Pi,U}(x)$ is the frequency of the category x in the corpus obtained by replacing each morpheme by its nearest ancestor category that is an element of G[2].

Here, $p_{\mathrm{L}}^{\Pi,U,G}(x;z)$ is the conditional probability that category x precedes category z, whereas $p_{\mathrm{R}}^{\Pi,U,G}(x;z)$ is the conditional probability that category x follows category z. We call this pair of probability distributions *the environment of category* z.

3 Generating a Category Set

3.1 Outline of Generating Algorithm

At the start of the algorithm, we set the category set to be a singleton containing only the root node. Then, we subdivide the category that must be subdivided, topdown.

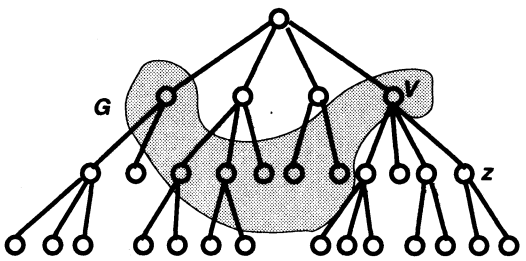

Figure 1: Category set before subdivision of category v

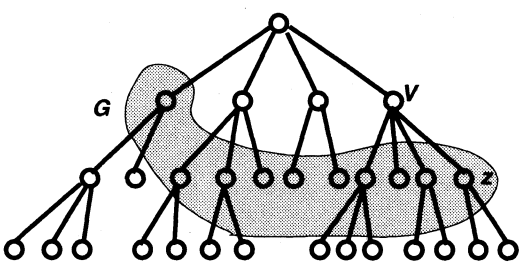

Figure 2: Category set after subdivision of category v

3.2 Subdivision Score

There are several possible methods of deciding whether we must subdivide a certain category or not. In this paper, we use a subdivision score based on Kullback-Leibler distance divergence.

First, we quantify the difference, with respect to environments, between category z and its parent category $v = P(z)$ by the following two terms:

$$D_{\mathrm{L}}^{\Pi,U,G}(z) \equiv \sum_{x \in G} p_{\mathrm{L}}^{\Pi,U,G}(x;z) \log \frac{p_{\mathrm{L}}^{\Pi,U,G}(x;z)}{\overline{p}_{\mathrm{L}}^{\Pi,U,G}(x;z)}$$

$$D_{\mathrm{R}}^{\Pi,U,G}(z) \equiv \sum_{x \in G} p_{\mathrm{R}}^{\Pi,U,G}(x;z) \log \frac{p_{\mathrm{R}}^{\Pi,U,G}(x;z)}{\overline{p}_{\mathrm{R}}^{\Pi,U,G}(x;z)},$$

where $\sum$ stands for summation over those instances in which the divisor is not zero. We express these terms simply as $D_{\mathrm{L}}^{v}(z)$ and $p_{\mathrm{L}}^{v}(x;z)$, if the suffixes are clear from the context.

Here, $\overline{p}_{\mathrm{L}}^{\Pi,U,G}(x;z)$ is the probability distribution that a category x precedes the parent category v. Based on the requirement on the category set described in Section 2.3, the parent category v

[2] In other words, the rewritten corpus is obtained by replacing each leaf category x by $\delta^{U,G}(x)$.

is not an element of G. It means $f_G^{\Pi,U}(v) = 0$. Therefore, we use the summation of the frequency of the sibling category $y \in C(v)$ in place of the frequency of the parent category v.

$$\overline{p}_{\mathrm{L}}^{\Pi,U,G}(x;z) \equiv \frac{\sum_{y \in C(P(z))} f_G^{\Pi,U}(x \cdot y)}{\sum_{y \in C(P(z))} f_G^{\Pi,U}(y)}$$

$$\overline{p}_{\mathrm{R}}^{\Pi,U,G}(x;z) \equiv \frac{\sum_{y \in C(P(z))} f_G^{\Pi,U}(y \cdot x)}{\sum_{y \in C(P(z))} f_G^{\Pi,U}(y)}$$

Now we define the difference, with respect to environments, between the category z and its parent category v, namely $D^{\Pi,U,G}(z)$, as the larger one of $D_{\mathrm{L}}^{\Pi,U,G}(z)$ and $D_{\mathrm{R}}^{\Pi,U,G}(z)$. This is because if the environments on either side of category z differ sufficiently from the parent's environment, the category z and its parent category v are different with respect to these environments.

Furthermore, we assign a value[3] of -1 to $D^{\Pi,U,G}(z)$, as shown below, when the frequency of category z is less than a certain threshold f_{T}. This is because we cannot judge whether the category must be subdivided or not when the frequency is very low.

$$D''(z) \equiv \begin{cases} -1 & , f_G^{\Pi,U}(z) < f_{\mathrm{T}} \\ \max(D''_{\mathrm{L}}(z), D''_{\mathrm{R}}(z)) & , \text{otherwise} \end{cases}$$

Finally, we define a subdivision score $E^{\Pi,U,G}(v)$, which indicates how significant it is to subdivide category v. If at least one child category is sufficiently different from the parent category, then the parent category may be subdivided. So, we define the subdivision score $E^{\Pi,U,G}(v)$ as the largest difference from among the differences between category v and each of its children.

$$E^{\Pi,U,G}(v) \equiv \max_{z \in C(v)} D^{\Pi,U,G \cup C(v) - \{v\}}(z)$$

It means that the subdivision score of a category v is the difference of the child which is most different from v. The subdivision score $E''(v)$ is 0, when a category v must not be subdivided. If a category v should be subdivided, $E''(v)$ has a larger value. And, the subdivision score $E''(v)$ has a value of -1, when we can not judge whether a category v must be subdivided or not.

3.3 Algorithm

Here, we summarize the algorithm described above.

(1) $G^0 \leftarrow \{\rho\}, t \leftarrow 1$.

(2) $\tilde{v}^t \leftarrow \underset{v \in G^{t-1} - V_{\mathrm{L}}(U)}{\operatorname{argmax}} E^{\Pi,U,G^{t-1}}(v)$.

(3) $\tilde{E}^t \leftarrow E^{\Pi,U,G^{t-1}}(\tilde{v}^t)$.

(4) if $\tilde{E}^t < E_{\mathrm{T}}$ then quit.

(5) $G^t \leftarrow G^{t-1} \cup C(\tilde{v}^t) - \{\tilde{v}^t\}$.

(6) $t \leftarrow t + 1$ and goto (2).

E_{T} is the subdivision score threshold used to decide when the algorithm must be terminated. Unsimilar subcategories are gathered into a category, if E_{T} is too low.

[3]This value, which indicates that the reliability of the category is low, need not be -1. We aim to distinguish the value from a divergence, which is nonnegative.

4 Experiments

We used RWC newspaper corpus tagged with the hierarchical part-of-speech system. The part-of-speech system had 509 leaf categories and 208 middle categories (or $|V_{\mathrm{M}}| = 509$ and $|V_{\mathrm{P}}| = 208$). We used the whole corpus containing 18,672 sentences, called "whole corpus", and 10 of its subcorpora. Table 1 shows the size of each of these corpora used in the experiment.

Table 1: Size of the corpora used in the experiment

name	Number of Sentences	Number of Morphemes
s01	934	44544
s02	867	42915
s03	853	44398
s04	901	44273
s05	891	45058
s06	946	44586
s07	888	44418
s08	1015	43811
s09	874	44121
s10	885	43896
whole	18672	888000

We set the frequency threshold f_{T} that specifies the significance of the subdivision score as 100. We also set E_{T} to be 0, because the proper value for E_{T} is unknown and we would like to estimate the proper value with this experiment.

Figure 3 shows the change in the subdivision score of the subdivided categories with time. The individual subdivided categories are shown in Tables 2, 3 and 4. Table 2 shows which

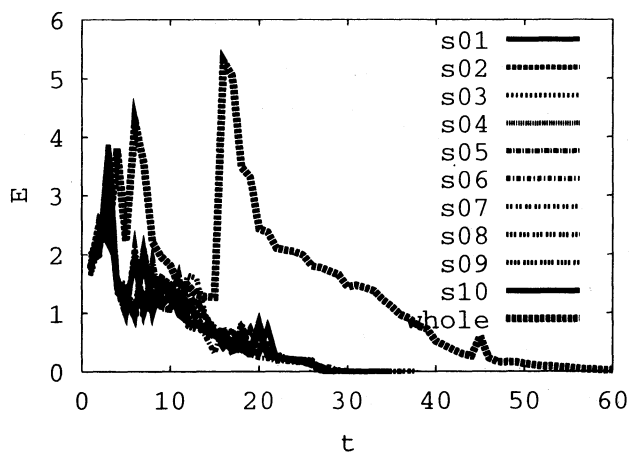

Figure 3: Subdivision score

categories are subdivided with the subcorpus "s01", and Tables 3 and 4 show which categories are subdivided with "whole corpus". In in Tables 2, 3 and 4, the third column labeled $\tilde{v}^t$ indicates the name of the subdivided category at time t. A hyphen stands for a parent-child relationship in their category names.

Figure 4 shows the change in the size of the category set G with time.

Table 2: Subdivided categories with subcorpus "s01"

t	$\tilde{E}^t$	$\tilde{v}^t$
001	1.876	ρ
002	1.984	ρ-JOSHI
003	3.867	ρ-N
004	1.597	ρ-JOSHI-KAKU JOSHI
005	1.050	ρ-V
006	1.915	ρ-V-SAHEN SURU
007	0.992	ρ-V-ICHIDAN
008	1.628	ρ-V-ICHIDAN-RENYOU TA SETSUZOKU
009	1.485	ρ-AUX
010	1.405	ρ-AUX-TOKUSHU
011	0.963	ρ-PREFIX
012	0.855	ρ-V-ICHIDAN-MIDASHI
013	0.812	ρ-V-GODAN RAGYOU
014	0.674	ρ-N-SUFFIX
015	0.611	ρ-N-FUKUSHI KANOU
016	0.528	ρ-V-GODAN WAGYOU SOKUONBIN
017	0.388	ρ-N-PN
018	0.528	ρ-N-PN-PERSON
019	0.341	ρ-V-ICHIDAN-RENYOU TAI SETSUZOKU
020	0.305	ρ-N-FUKUSHI HI JIRITSU
021	0.758	ρ-N-FUKUSHI HI JIRITSU-FUKUSHI KANOU
022	0.235	ρ-ADJ
023	0.226	ρ-AUX-KEIYOUSHI GATA
024	0.218	ρ-N-PN-RG
025	0.203	ρ-ADV
026	0.140	ρ-ADJ-MIDASHI
027	0.019	ρ-V-GODAN WAGYOU SOKUONBIN-MIDASHI
028	0.009	ρ-V-GODAN RAGYOU-RENYOU TA SETSUZOKU
029	0.004	ρ-N-SUFFIX-JOSUSHI
030	0.000	ρ-V-GODAN RAGYOU-MIDASHI

Table 3: Subdivided categories with whole corpus (1)

t	$\tilde{E}^t$	$\tilde{v}^t$
001	1.768	ρ
002	2.527	ρ-ADJ
003	2.230	ρ-AUX
004	3.800	ρ-AUX-TOKUSHU
005	2.227	ρ-N
006	4.259	ρ-N-SUFFIX
007	3.554	ρ-JOSHI
008	2.212	ρ-AUX-KEIYOUSHI GATA
009	1.959	ρ-AUX-DOUSHI GATA
010	1.816	ρ-ADJ-MIDASHI
011	1.488	ρ-N-SUFFIX-JOSUSHI
012	1.478	ρ-JOSHI-KAKU JOSHI
013	1.317	ρ-N-FUKUSHI HI JIRITSU
014	1.276	ρ-PREFIX
015	1.274	ρ-V
016	5.294	ρ-V-ICHIDAN
017	5.036	ρ-V-SAHEN SURU
018	3.485	ρ-V-GODAN WAGYOU SOKUONBIN
019	3.310	ρ-V-GODAN RAGYOU
020	2.434	ρ-V-GODAN KAGYOU IONBIN
021	2.375	ρ-V-GODAN WAGYOU SOKUONBIN-MIDASHI
022	2.108	ρ-V-GODAN SAGYOU
023	2.073	ρ-V-GODAN WAGYOU SOKUONBIN-RENYOU TA SETSUZOKU
024	2.052	ρ-V-GODAN MAGYOU
025	1.986	ρ-V-ICHIDAN-MIZEN NAI SETSUZOKU
026	1.800	ρ-V-ICHIDAN-MIDASHI
027	1.773	ρ-V-ICHIDAN-RENYOU TA SETSUZOKU
028	1.714	ρ-V-GODAN BAGYOU
029	1.649	ρ-V-ICHIDAN-RENYOU TAI SETSUZOKU
030	1.445	ρ-V-KAHEN

Table 4: Subdivided categories with whole corpus (2)

t	$\tilde{E}^t$	$\tilde{v}^t$
031	1.470	ρ-V-KAHEN-RENYOU TA SETSUZOKU
032	1.430	ρ-V-ICHIDAN-RENYOU TAI SETSUZOKU
033	1.373	ρ-V-GODAN KAGYOU SOKUONBIN
034	1.224	ρ-V-GODAN TAGYOU
035	1.111	ρ-N-SUFFIX-FUKUSHI
036	0.970	ρ-V-GODAN GAGYOU
037	0.905	ρ-V-GODAN RAGYOU-MIZEN NAI SETSUZOKU
038	0.779	ρ-V-ICHIDAN-RENYOU MASU SETSUZOKU
039	0.726	ρ-N-FUKUSHI HI JIRITSU-FUKUSHI KANOU
040	0.518	ρ-N-FUKUSHI HI JIRITSU
041	0.436	ρ-JOSHI-FUKU/HEIRETSU/SHU
042	0.364	ρ-V-KAHEN-MIDASHI
043	0.297	ρ-V-GODAN KAGYOU SOKUONBIN-MIDASHI.
044	0.260	ρ-N-PN
045	0.598	ρ-N-PN-PERSON
046	0.229	ρ-V-ICHIDAN-MIZEN NU SETSUZOKU
047	0.164	ρ-N-PN-RG
048	0.159	ρ-V-SAHEN X SURU
049	0.158	ρ-V-ICHIDAN-MIZEN NU SETSUZOKU
050	0.130	ρ-ADV
051	0.111	ρ-ADJ-RENYOU TA SETSUZOKU
052	0.091	ρ-V-GODAN WAGYOU SOKUONBIN-RENYOU TAI SETSUZOKU
053	0.089	ρ-V-SAHEN SURU-RENYOU MASU SETSUZOKU
054	0.074	ρ-AUX-BUNGO
055	0.071	ρ-V-GODAN WAGYOU SOKUONBIN-RENYOU MASU SETSUZOKU
056	0.049	ρ-ADJ-RENYOU TE SETSUZOKU
057	0.034	ρ-V-GODAN KAGYOU IONBIN-RENYOU TAI SETSUZOKU
058	0.033	ρ-V-ICHIDAN-MIZEN RERU SETSUZOKU
059	0.025	ρ-V-GODAN SAGYOU-GODAN SAGYOU
060	0.021	ρ-V-GODAN RAGYOU-RENYOU MASU SETSUZOKU

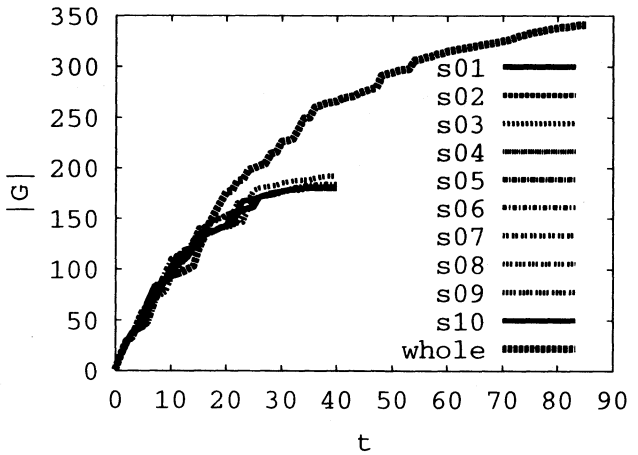

Figure 4: Size of category set

5 Discussion

5.1 Adequacy of Subdivision

Figure 3 shows that all the subcorpora are almost equivalent in the transition of the maximum subdivision score. It also means that, if we use such subcorpora, the generated category set has little noise.

However, comparing with the results from the whole corpus, we notice some differences. As Table 2, Table 3 and Table 4 show, some categories that are subdivided early with the whole corpus are subdivided later in case of subcorpora. For example, ρ-ADJ[4], which is subdivided second with the whole corpus, is subdivided at $t = 22$ with the s01 corpus. We conjecture that the category corresponding to adjectives is subdivided later because of the low frequency of the adjectives. A category has a low subdivision score if its children categories do not occur uniformly, although the frequency of the category may be high.

Once a certain category is subdivided, the subdivision scores of its children categories tend to be larger than the subdivision scores of its sibling categories. For example, in Table 2, because the category ρ-V is subdivided early, its children categories are subdivided in succeeding steps. In the part-of-speech system we use, the category ρ-V has many descendant categories. The number of leaf categories which are descendant of ρ-V is 361. It is more than half of 509, which is the total number of leaf categories. As a result, category ρ-ADJ is subdivided later. In fact, when the category corresponding to verbs were subdivided, the maximum subdivision shot up, as shown in Figure 3 and Table 3.

5.2 Proper Value for the Threshold of Subdivision Score

It is difficult to estimate a proper value for the threshold of the subdivision score E_{T}, because the maximum subdivision score often shoots up.

If we set $E_{\mathrm{T}} = 2$, the algorithm terminates at $t = 8, |G| = 83$ with the whole corpus. But the largest subdivision score appears at $t = 16$. If we set $E_{\mathrm{T}} = 1.2$ in order to call the largest subdivision score, the algorithm terminates at $t = 34, |G| = 249$. The category set at $t = 34$ seems to be too large.

[4] ADJ and V stand for adjective and verb, respectively.

5.3 Future Work

In order to solve the problem mentioned in Section 5.2, we are considering improving the generating algorithm. We are considering other methods for separating a category from its parent instead of subdividing a category at each step. Furthermore, we are considering using other part-of-speech systems in which the leaf nodes are morphemes and not parts-of-speech.

In this work, we aimed to generate a proper category set. But this work lacks an objective evaluation of the aptness of a category set. We have to find a measure of aptness and evaluate the algorithm by this measure. We would like to use a linguistic one.

Using this measure, we have to run experiments that determine two parameters of the generating algorithm. Then, we have to compare our method to other existing methods.

We think we have to perform more experiments with other data. For example, we would like to do an experiment using a larger corpus or an experiment with corpus in other domains or other languages.

6 Conclusion

In this paper, we proposed a method to generate a proper categorization of morphemes given a hierarchical part-of-speech system and a corpus tagged using this part-of-speech system. Then, we ran experiments to test the generating algorithm. As a result, we obtained a moderately proper categorization, and found several candidates for improvement.

Acknowledgements

We would like to thank Bipin Indurkhya for his advice on expression in English.

References

Brants, Thorsten. 1995. Tagset reduction without information loss. In *Meeting of the Association for Computational Linguistics*, pages 287–289.

Déjean, Hervé. 2000. How to evaluate and compare tagsets: a proposal. In *LREC 2000*.