Crosslingual Countability Classification: English meets Dutch

Leonoor van der Beek
Alfa-informatica
RuG, Pb 716
9700 AS Groningen
The Netherlands
vdbeek@let.rug.nl

Timothy Baldwin
CSLI
Stanford University
Stanford, CA 94305 USA
tbaldwin@csli.stanford.edu

Abstract

This paper presents a range of methods for classifying Dutch nouns as countable, uncountable or plural only based on both Dutch and English data. The classification is based on the occurrence of countability specific linguistic features that are extracted from unannotated corpora. We show that in the absence of reliable Dutch gold standard data, cross-linguistic classification can be achieved on the basis of a word-to-word or feature-to-feature mapping between English and Dutch.

1 Introduction

This paper presents several methods for learning countability preferences of Dutch nouns using both Dutch and English data. Knowledge of countability is important both for analysis and generation. In analysis it helps to constrain the set of possible parses and their interpretation. A generator uses countability information to determine whether a noun can become plural and what determiners it can combine with.

In the case of English, there exist high-quality sources of countability data (see § 2.4). This makes it possible to use the standard supervised classification paradigm in utilising a suitable feature space to classify unannotated nouns based on the gold standard data (Baldwin and Bond, 2003a; Baldwin and Bond, 2003b). For Dutch, however, there is no reliable source of countability information, leading us to consider the crosslingual approach described in this paper.

The assumption underlying this work is that Dutch and English are sufficiently close linguistically that there is a strong correlation between countabilities of nouns in the two languages. Both languages distinguish countable, uncountable and plural only nouns.[1] Although mismatches exist—e.g. *hersenen* (plural only) vs. *brain* (countable), *onweer* (uncountable) vs. *thunderstorm* (countable)—many Dutch words are in the same countability class as their English equivalents (e.g. *auto/car, eten/food*, *goederen/goods*). One obvious approach, therefore, is to simply map the countabilities of English nouns onto their Dutch counterparts.

A less direct approach to crosslingual countability transfer is to base classification on corpus occurrence with linguistic predictors of the different countability classes. Linguistic features that are associated with the countability classes often have direct translations in the other language (e.g. syntactic number, co-occurrence with denumerators) or can be mapped onto an equivalent feature (e.g. the English N_1 *of* N_2 construction and Dutch measure noun construction—see § 2.2). In some cases however, the mapping is imperfect (e.g. *much* occurs only with uncountable nouns, but the Dutch translation *veel* is also the translation of

[1] A fourth class of bipartite nouns (e.g. *scissors, trousers*) is generally recognised for English, but has no Dutch correlate.

many, and occurs with both uncountable singular and countable plural nouns) or no equivalent exists in one of the languages (e.g. the occurrence of a plural noun as a modifier indicates plural only in English, but not in Dutch).

2 Preliminaries

In this section, we describe the countability classes, the feature space, two distinct feature abstractions, the feature extraction method and resources used in this research.

2.1 Countability classes

Dutch and English nouns are classified as belonging to one or more of three possible classes: countable, uncountable and plural only. **Countable** nouns can be modified by denumerators, prototypically numbers, and have a morphologically marked plural form: *een hond/one dog*, *twee honden/two dogs*. **Uncountable** nouns cannot be modified by denumerators, but can be modified by unspecific quantifiers such as *veel/much*, and do not show any number distinction (prototypically being singular) **een rijst/one rice*, *een beetje rijst/some rice*, **twee rijsten/two rices*. This class includes many abstract nouns, material denoting nouns, generics and deverbalised nouns. **Plural only** nouns only have a plural form, such as *goederen/goods* and cannot be denumerated. Many plural only nouns, such as *kleren/clothes*, use the plural form even as modifiers (*klerenkast* "clothes-closet" and *a clothes horse*) even though in English, only bare nouns appear as modifiers. The plural only class is by far the smallest set of nouns. At least in Dutch, it is considered to be a closed class, making the automated classification largely superfluous. Note that the distinctions are in fact not categorical: prototypical count nouns can be used in an uncountable context, forcing a 'substance' interpretation (the **universal grinder**) and uncountable nouns can in certain contexts be denumerated, resulting in a 'unit' interpretation (the **universal packager**) (Allan, 1980). However, this does not contradict our assumption that nouns have a basic classification as countable or uncountable.

2.2 Feature space

We use a basic feature space to formulate two feature abstractions in this research, based on analysis of (a) the overall occurrence of each type and (b) token-based occurrence in particular constructions which are indicative of a given countability class.

The feature space is made up of **feature clusters**, each of which is conditioned on the occurrence of a **target noun** in a given construction. Feature clusters are either one-dimensional (describe a single multivariate feature) or two-dimensional (describe the interaction between two multivariate features), with each dimension describing a lexical or syntactic property of the construction in question. Below, we provide a basic description of the 9 feature clusters used in this research and their dimensionality ($^{[x]}\mathbf{x}$=1-dimensional feature cluster with x unit features for language X, $^{[x\times y]}\mathbf{x}$= 2-dimensional feature cluster with $x \times y$ unit features for language X). For further details and predicted correlations between feature values and particular countability classes for English, the reader is referred to Baldwin and Bond (2003a)

Head noun number: $^{[2]}\mathbf{E}$ *vs.* $^{[2]}\mathbf{D}$ the number of the target noun when it heads an NP

Subject–verb agreement: $^{[2\times2]}\mathbf{E}$ *vs.* $^{[2\times2]}\mathbf{D}$ the number of the target noun in a subject position vs. number agreement on the governing verb

Coordinate noun number: $^{[2\times2]}\mathbf{E}$ *vs.* $^{[2\times2]}\mathbf{D}$ the number of the target noun vs. the number of the head nouns of conjuncts

N_1 *of* N_2/measure noun constructions: $^{[11\times2]}\mathbf{E}$ *vs.* $^{[11\times2]}\mathbf{D}$ the type of the N_1 vs. the number of the target noun (N_2) in an English N_1 *of* N_2 construction (e.g. *a group of people*) or Dutch measure noun construction (e.g. *een groep mensen*). We have identified a total of 11 N_1 types for use in this feature cluster (e.g. `COLLECTIVE`, `LACK`, `TEMPORAL`).

Occurrence in PPs: $^{[52\times2]}\mathbf{E}$ *vs.* $^{[84\times2]}\mathbf{D}$ the preposition type and presence or absence of

a determiner when the target noun occurs in **singular** form in a PP.

Pronoun co-occurrence: $[12\times2]_{\mathbf{E}}$ *vs.* $[7\times2]_{\mathbf{D}}$ what personal, reflexive and possessive pronouns occur in the same sentence as singular and plural instances of the target noun.

Singular determiners: $[10]_{\mathbf{E}}$ *vs.* $[10]_{\mathbf{D}}$ what singular-selecting determiners occur in NPs headed by the target noun in **singular** form.

Plural determiners: $[12]_{\mathbf{E}}$ *vs.* $[13]_{\mathbf{D}}$ what plural-selecting determiners occur in NPs headed by the target noun in **plural** form.

Number-neutral determiners: $[11\times2]_{\mathbf{E}}$ *vs.* $[13\times2]_{\mathbf{D}}$ what number-neutral determiners occur in NPs headed by the target noun, and what is the number of the target noun for each.

The Dutch and English feature clusters represent the same linguistic structures, even if the individual features are not direct translations of each other. The only exception is the N_1 *of* N_2/measure noun construction where markedly different constructions in the two languages express the same concept (a quantity of something) and bring about the same restrictions with respect to countability.

2.3 Feature extraction

We use a variety of pre-processors to map the raw data onto the types of constructions targeted in the feature clusters, namely a POS tagger and a full-text chunker for both English and Dutch, and additionally a dependency parser for English. For Dutch, pos-tags, lemma's and chunk data were extracted from automatically generated, fully-parsed Alpino output. For English, we used an fnTBL-based tagger (Ngai and Florian, 2001) with the Penn tagset, morph (Minnen et al., 2001) as our lemmatiser, an fnTBL-based chunker which runs over the output of the tagger, and RASP (Briscoe and Carroll, 2002) as the dependency parser.

These data sets are then used independently to test the efficacy of the different systems at capturing features used in the classification process, or in tandem to consolidate the strengths of the individual methods and reduce system-specific idiosyncrasies in the feature values. When combining the Dutch and English in classification, we invariably combine like systems (e.g. Dutch tagger-derived data with English tagger-derived data).

The English data was extracted from the written component of the British National Corpus (Burnard, 2000), and the Dutch data from the newspaper component of the Twente Nieuws Corpus.[2]

After generating the different feature vectors for each noun based on the above configurations, we filtered out all nouns which did not occur at least 10 times in NP head position according to the output of all pre-processors. This resulted in 20,530 English nouns and 12,734 Dutch nouns.

2.3.1 Distribution-based classification

In distribution-based classification, we take each target noun in turn and compare its amalgamated value for each unit feature with (a) the values for other target nouns, and (b) the value of other unit features within that same feature cluster (Baldwin and Bond, 2003b).

In the case of a one-dimensional feature cluster, each unit feature f_s for target noun w is translated into the 3-tuple:

$$\langle \frac{freq(f_s|w)}{freq(f_s|*)}, \frac{freq(f_s|w)}{freq(w)}, \frac{freq(f_s|w)}{\sum_i freq(f_i|w)} \rangle$$

In addition to mapping individual unit features onto 3-tuples, we introduce a 3-tuple for each feature cluster representing the sum over all member values.

In the case of a two-dimensional feature matrix, each unit feature $f_{s,t}$ for target noun w is translated into the 5-tuple:

$$\langle \frac{freq(f_{s,t}|w)}{freq(f_{s,t}|*)}, \frac{freq(f_{s,t}|w)}{freq(w)}, \frac{freq(f_{s,t}|w)}{\sum_{i,j} freq(f_{i,j}|w)}, \frac{freq(f_{s,t}|w)}{\sum_i freq(f_{i,t}|w)}, \frac{freq(f_{s,t}|w)}{\sum_j freq(f_{s,j}|w)} \rangle$$

[2] `http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html`

As for one-dimensional feature clusters, we introduce amalgamated features for each row and column of the feature matrix, and describe each in the form of a one-dimensional 3-tuple. For further details, see the description of the monolingual English task in Baldwin and Bond (2003a).

The feature clusters produce a combined total of 1284 individual feature values for English and 1664 feature values for Dutch.

2.3.2 Agreement-based classification

In the case of noun countability, extensive linguistic literature exists documenting diagnostics for the different classes both for English (Quirk et al., 1985; Huddleston and Pullum, 2002) and for Dutch (Haeseryn et al., 1997). It is possible to use this to identify the features which are positively-correlated with a unique countability class, and to determine the combined token-level agreement for each countability class and pre-processor system pairing. The number of diagnostics considered for each of the countability classes is: 32 for English and 11 for Dutch countable nouns, and 18 for English and 5 for Dutch uncountable nouns. No reliable single-feature diagnostics exist for plural only nouns in Dutch, so this class was excluded from agreement-based classification.

The token-level correlation for each feature f_s is calculated fourfold according to relative agreement, the κ statistic, correlated frequency and correlated weight, each of which is calculated over the total diagnostic evidence for each class instance of a given target noun. The **relative agreement** between systems s_1 and s_2 wrt f_s for target noun w is defined to be:

$$agree_{(f_s,w)}(s_1, s_2) = \frac{|token_{(f_s,w)}(s_1) \cap token_{(f_s,w)}(s_2)|}{|token_{(f_s,w)}(s_1) \cup token_{(f_s,w)}(s_2)|}$$

where $token_{(f_s,w)}(s_i)$ returns the set of token instances of f_s. The κ **statistic** (Carletta, 1996) is modified slightly from its original form to:

$$\kappa_{f_s}(s_1, s_2) = \frac{agree_{(f_s,w)}(s_1, s_2) - \frac{\sum agree_{(f_s,*)}(s_1,s_2)}{N}}{1 - \frac{\sum agree_{(f_s,*)}(s_1,s_2)}{N}}$$

That is, it represents the divergence in relative agreement wrt f_s for target noun w, relative to the mean relative agreement wrt f_s over all words. **Correlated frequency** is defined to be:

$$cfreq_{(f_s,w)}(s_1, s_2) = \frac{|token_{(f_s,w)}(s_1) \cap token_{(f_s,w)}(s_2)|}{freq(w)}$$

That is, it describes the occurrence of tokens in agreement for f_s relative to the total occurrence of the target word.

We additionally calculate the overall **correlated weight** for each countability class C as:

$$cw_{(C,w)}(s_1, s_2) = \frac{\sum_{f_s \in C} |token_{(f_s,w)}(s_1) \cap token_{(f_s,w)}(s_2)|}{\sum_i |token_{(f_i,w)}(s_1) \cap token_{(f_i,w)}(s_2)|}$$

Correlated weight describes the occurrence of correlated features in the given countability class relative to other correlated features.

We initially derive a 3-tuple for each of the pre-processor system pairings, and convert this into a single 3-tuple by averaging over the system pairings.

2.4 Gold standard data

Information about English noun countability was obtained from two sources: **COMLEX** 3.0 (Grishman et al., 1998) and the common noun part of **ALT-J/E**'s Japanese-to-English semantic transfer dictionary. We generated the positive exemplars for the countable and uncountable classes from the intersection of the **COMLEX** and **ALT-J/E** data for that class; negative exemplars, on the other hand, are those not annotated as belonging to that class in either lexicon. With the plural only data, **COMLEX** cannot be used as it does not describe this class. We thus took all members of each class listed in **ALT-J/E** as our positive exemplars, and all remaining nouns with non-identical singular and plural forms as negative exemplars. This resulted in the following datasets:

Class	***+ve data***	***−ve data***
Countable	4342	1476
Uncountable	1519	5471
Plural only	84	5639

In Dutch, there are two electronic dictionaries with countability information: CELEX (Baayen et al., 1993) and the lexicon of Alpino (Bouma et al., 2000). The latter includes the former as well as the Parole lexicon (no countability information) and is manually modified and extended.

We thus used the Alpino data to generate the training data for the monolingual Dutch classifiers, classifying lexical entries that do not have a singular form as plural only. This resulted in the following datasets:

Class	*+ve data*	*−ve data*
Countable	11570	2857
Uncountable	2867	11560
Plural only	27	14400

In order to both evaluate the various classifiers and gauge the reliability of the Alpino countability judgements, we manually annotated 207 nouns based on the Twente Nieuws Corpus data. The agreement[3] in countability judgements between the Alpino lexicon and hand-annotated data is a modest 73.1%, underlining the brittleness of the Alpino data.

3 Classifier design

We propose a variety of both monolingual and crosslingual unsupervised and supervised classifier architectures for the task of learning countability. A separate classifier is built for each countability class, rather than having a common classifier for all class combinations. This is based on the results of thorough evaluation of these two architectures over monolingual data, and the finding that the classifier suite architecture is superior (Baldwin and Bond, 2003b). In all cases, our classifiers are built using TiMBL version 4.2 (Daelemans et al., 2002), a memory-based classification system based on the k-nearest neighbour algorithm. TiMBL was used with the default configuration except that k was set to 9 throughout.

3.1 Monolingual unsupervised classifiers

In an attempt to derive a baseline for each countability class/pre-processor system combination, we built a monolingual unsupervised classifier. For each target noun, the unsupervised classifier simply checks for the existence of diagnostic data (as used in agreement-based classification) in the output of each of the POS tagger and chunker for the given countability class (*Unsupervised(POS)* and *Unsupervised(chunk)*, respectively). We perform basic system combination by voting between the two pre-processor datasets as to whether the target noun belongs to a given countability class, and breaking ties in favour of the majority class (*Unsupervised(all)*).

3.2 Monolingual supervised classifiers

Despite our reservations about the quality of countability annotation in the Alpino lexicon, we implemented a conventional monolingual classifier over the full 1664-feature distribution data for Dutch. We trained our classifiers over the Alpino data as detailed in § 2.4,[4] using first the feature values from each of the POS tagger and chunker (*NN(POS)* and *NN(chunk)*, respectively), and finally the feature values averaged across the tagger and chunker (*NN(all)*).

3.3 Crosslingual classifiers

Below, we describe each of the crosslingual (supervised) classifier architectures.

Translation-based classification

Translation-based classification applies the observation made in § 1 that Dutch nouns often take the same countability as their English translation equivalents. We thus use a monolingual supervised classifier—as described above for Dutch—to learn countabilities in English, extract translation pairs from a bilingual dictionary, and transfer the English countabilities directly across to their Dutch counterparts. In the case of multiple English translations for a given Dutch noun, we vote between them to produce a single judgement for each countability class (*EN(translation)*). All results are based on a single bilingual dictionary, English–Dutch freedict version 1.1-1, of 15,426 Dutch entries.

Cluster-to-cluster classification

As observed above (§ 2.2), there is a strong correlation between the feature clusters used for Dutch and English. The most straightforward way of applying this correlation in a crosslingual classifier is to align the 3-tuple totals for

[3] I.e. the proportion of word-level countability class assignments over which the two systems agreed.

[4] Note that the set of 207 hand-annotated nouns are held out from the Alpino training data

each one-dimensional feature cluster and match 3-tuple totals for each two-dimensional feature cluster (e.g. for the PP feature, we align the totals for the singular and plural features but not the totals for each individual preposition), and ignore all values for individual unit features. The total number of aligned features is 52, and cluster-to-cluster classification is carried out over the output of the POS tagger, chunker and combined outputs of the two pre-processors (*EN(cluster,∗)*).

Feature-to-feature classification

Feature-to-feature classification utilises the same feature alignments as cluster-to-cluster classification, but also aligns individual unit features wherever possible. In the case of singular determiners, for example, the English *a*, *a certain* and *one* align with the Dutch *een*, *ene* and *1*, respectively. The remainder of Dutch determiners have no no one-to-one correspondence with an English determiner but select for (singular) countable nouns. They are thus aligned with all remaining English determiners which similarly select for (singular) countable nouns in a many-to-many fashion, by averaging over the individual values of each 3-tuple. This accounts for members of the Dutch feature cluster but leaves the set of English determiners which select for uncountable nouns (i.e. *little*, *much*[5] and *much more*); these are pruned from the feature space. Feature-to-feature classification results in a total of 640 features and is tested over the output of the POS tagger, chunker and combined outputs of the two pre-processors (*EN(feature,∗)*).

Agreement-based classification

Agreement-based classification involves computing the 3-way agreement statistics for each feature positively-correlated with the countable and uncountable classes,[6] and taking the mean of each statistic for each of the classes; we additionally calculate the correlated weight for each class. A total of 8 features are generated (*EN(agree,all)*).

[5] *Much* aligns with the Dutch *veel*, which occurs in the number-neutral feature cluster as it selects for singular uncountable nouns and plural countable nouns. Here, we make no attempt to align between disparate feature clusters due to the fundamental semantic differences between them.

[6] Recall that we do not carry out agreement-based classification for plural only nouns due to a lack of reliable diagnostics.

Method	***Acc***	***P***	***R***	***F***
Majority class	.802	.802	1.000	.890
Unsupervised(POS)	.502	.970	.392	.558
Unsupervised(chunk)	.527	.947	.434	.595
Unsupervised(all)	.565	.942	.488	.643
NN(POS)	.785	.891	.837	.863
NN(chunk)	.780	.885	.837	.861
EN(translation)	**.820**	.900	**.882**	**.891**
EN(cluster,POS)	.785	.918	.807	.859
EN(cluster,chunk)	.790	.913	.819	.863
EN(cluster,all)	.737	.924	.735	.819
EN(feature,POS)	.620	.949	.560	.705
EN(feature,chunk)	.668	.938	.633	.755
EN(feature,all)	.561	.942	.488	.643
EN(agree,all)	.425	.956	.297	.453
Combined	.811	.892	.874	.883
EE	—	.948	.972	.960

Table 1: Results for countable nouns

3.4 System combination

System combination involves taking the outputs of heterogeneous classifiers and making a consolidated classification based upon them. It has been shown to be effective in tasks ranging from word sense disambiguation to tagging in bettering the performance of component systems (Klein et al., 2002; van Halteren et al., 2001). In our case, we take the outputs of all classifiers (excluding the baseline majority class classifier, for a total of 13 classifiers) for each countability class, and run TiMBL over them (*Combined*).

4 Results and Discussion

Classifier performance is rated according to classification accuracy (the proportion of instances classified correctly: ***Acc***), precision (***P***), recall (***R***) and F-score ($\beta = 1$: ***F***). The label *NaN* for precision, recall and F-score represents a value of $\frac{0}{0}$.

We present the results for four monolingual baseline systems for each countability class: a majority-class classifier and an unsupervised classifier run over each of the tagger, chunker and combined system outputs.[7] The **major-**

[7] The results for the three unsupervised classifiers are

Method	*Acc*	*P*	*R*	*F*
Majority class	.657	(.343)	(1.000)	(.511)
Unsupervised(POS)	.623	.468	.718	.567
Unsupervised(chunk)	.507	.401	.887	.553
Unsupervised(all)	.614	.457	.676	.545
NN(POS)	.746	.694	.479	.567
NN(chunk)	.722	.646	.436	.521
EN(translation)	.758	.667	.235	.348
EN(cluster,POS)	.732	.643	.507	.567
EN(cluster,chunk)	**.766**	.735	.507	.600
EN(cluster,all)	.756	**.714**	.493	.583
EN(feature,POS)	.751	.647	.620	.633
EN(feature,chunk)	.761	.712	.521	.602
EN(feature,all)	.756	.624	.746	**.679**
EN(agree,all)	.575	.464	**.853**	.601
Combined	.801	.787	.637	.704
EE	—	.884	.907	.895

Table 2: Results for uncountable nouns

Method	*Acc*	*P*	*R*	*F*
Majority class	.957	(.043)	(1.000)	(.083)
Unsupervised(∗)	.957	*NaN*	.000	*NaN*
NN(POS)	.966	*NaN*	.000	*NaN*
NN(chunk)	.966	*NaN*	.000	*NaN*
EN(translation)	.952	*NaN*	.000	*NaN*
EN(cluster,POS)	**.971**	**1.000**	.143	.250
EN(cluster,chunk)	**.971**	.667	**.286**	**.400**
EN(cluster,all)	.961	.000	.000	*NaN*
EN(feature,POS)	**.971**	**1.000**	.143	.250
EN(feature,chunk)	.969	*NaN*	.000	*NaN*
EN(feature,all)	.966	.500	**.286**	.364
Combined	.957	*NaN*	.000	*NaN*
EE	—	.800	.457	.582

Table 3: Results for plural only nouns

ity class system simply classifies all instances according to the most commonly-attested class in the given dataset. Irrespective of the majority class, we calculate the F-score based on a positive-class classifier, i.e. a classifier which naively classifies each instance as belonging to the given class; in the case that the positive class is not the majority class, the F-score is given in parentheses.

The results for each classifier are presented in Tables 1–3, broken down into the three countability classes. In each table, the best single value for each of evaluation metrics (other than for the combined system) is presented in **boldface**. For each class, we also present the precision, recall and F-score for the best-performing classifier on the monolingual English countability classification task (based on the results in Baldwin and Bond (2003b): *EE*). This is intended to provide an upper bound for the task, assuming relatively noise-free training data and full feature correlation.

Perhaps the first thing to notice is how much better the classifiers perform for countable nouns than uncountable nouns, with plural only nouns producing very modest results. This is due to two factors: the relative occurrence of members of the three classes (as reflected in the majority class classification accuracies), and the relative volume of features correlated with each class. As mentioned above, plural only nouns are generally considered to be a closed class in Dutch, making the results for the plural only classifiers largely academic. The relatively high baseline accuracy and F-score for countable nouns (.802 and .890) surpassed the performance of all classifiers other than the translation-based and combined classifiers (*EN(translation)* and *Combined*, respectively). For uncountable nouns, on the other hand, appreciable gains over the baseline were observed for many of the systems.

We have made the claim that, due to the lack of reliable training data in Dutch, crosslingual classification of countability using English data is a more viable option. This is borne out by the result that, for all countability classes, the best of the simple (non-combined) supervised classifiers is a crosslingual system for all evaluation measures. The margin of crosslingual systems over monolingual systems is particularly great for uncountable and plural only nouns.

The highly commendable results for the translation-based classifier over the countable class require qualification. The presented results are based on those Dutch nouns for which the dictionary contained translations with learned countabilities, which total only 63 out of the 207 Dutch nouns. That is, assuming we have translation data and the English translations are countable, then there is a very high likelihood that the Dutch count is countable. Note that the translation-based classifier is much less successful over uncountable and plural only nouns.

The relative results for the different classifiers

identical for the plural only class, so they are presented as a single classifier.

over tagged, chunked and combined data are variable, although the chunker appears to do a slightly better job of extracting countable features, and conversely the tagger seems slightly more adept at extracting uncountable features. Combination of the pre-processor outputs has notable successes and failures.

System combination leads to a higher combined classification accuracy and F-score than any one component system in all cases. In order to test the relative quality of the output of the combined classifiers as compared to the original Alpino data, we calculated the overall class agreement relative to the gold standard countabilities, returning a figure of 71.1%. Recall that the agreement for the Alpino data was 73.1%, such that the Alpino data is only marginally more consistent with the corpus-attested countabilities. That is, our method allows us to automatically annotate countability to approximately the same level of accuracy as a mature computational lexicon.

5 Conclusion

We have presented several methods for classifying Dutch nouns as countable, uncountable or plural only on the basis of Dutch and English data. The classifiers crucially depend on the linguistic features that were extracted from unannotated corpora and that are associated with a particular countability class. In the absence of reliable gold standard data, we developed a set of cross-linguistic classifiers, mapping English countability preferences onto Dutch, both on the word level and on the feature level. With variable pre-processors and classification strategies, the mono-linguistic classifiers were always outperformed by at least one cross-linguistic classifier. The agreement with the gold standard data achieved by the combined classifier was 71.1%, which is comparable to the agreement for the Alpino lexicon.

Acknowledgements

The first author was supported by the National Science Foundation under Grant No. BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. The second author was supported by the NWO Council for the Humanities, the Graduate School for Behavioral and Cognitive Neurosciences (BCN) and the Center for Language and Cognition Groningen (CLCG). The research was carried out within the framework of the PIONIER Project *Algorithms for Linguistic Processing*, which is funded by NWO (Dutch Organization for Scientific Research) and the University of Groningen.

We would like to thank Francis Bond, Ann Copestake, Dan Flickinger, Gertjan van Noord, Ivan Sag and three anonymous reviewers for their valuable input on this research, and John Carroll for providing access to RASP.

References

Keith Allan. 1980. Nouns and countability. *Language*, 56(3):541–67.

Harald R. Baayen, Richard Piepenbrock, and Hedderik van Rijn. 1993. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Timothy Baldwin and Francis Bond. 2003a. Learning the countability of English nouns from corpus data. In *Proc. of the 41st Annual Meeting of the ACL*, Sapporo, Japan. (to appear).

Timothy Baldwin and Francis Bond. 2003b. A plethora of methods for learning English countability. In *Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, Sapporo, Japan. (to appear).

Gosse Bouma, Gertjan van Noord, and Rob Malouf. 2000. Alpino: Wide coverage computational analysis of Dutch. In *Computational Linguistics in the Netherlands (CLIN 2000)*.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21:543–65.

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499–1504, Las Palmas, Canary Islands.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

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

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2002. TiMBL: Tilburg memory based learner, version 4.2, reference guide. ILK technical report 02-01.

Ralph Grishman, Catherine Macleod, and Adam Myers, 1998. *COMLEX Syntax Reference Manual*. Proteus Project, NYU. (`http://nlp.cs.nyu.edu/comlex/refman.ps`).

W. Haeseryn, K. Romijn, G. Geerts, J. de Rooij, and M.C. van den Toorn, editors. 1997. *Algemene Nederlandse Spraakkunst*. Nijhoff, Groningen.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Dan Klein, Kristina Toutanova, H. Tolga Ilhan, Sepandar D. Kamvar, and Christopher D. Manning. 2002. Combining heterogeneous classifiers for word-sense disambiguation. In *Proc. of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Pittsburgh, USA.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–23.

Grace Ngai and Radu Florian. 2001. Transformation-based learning in the fast lane. In *Proc. of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2001)*, pages 40–7, Pittsburgh, USA.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London, UK.

Hans van Halteren, Jakub Zavrel, and Walter Daelemans. 2001. Improving accuracy in word class tagging through combination of machine learning systems. *Computational Linguistics*, 27(2):199–230.