In Proceedings of the 2003 Australasian Language Technology Workshop (ALTW2003),
Melbourne, Australia, pp. 33-40.

The Ins and Outs of Dutch Noun Countability Classification

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

Leonoor van der Beek
University of Groningen
Pb 716, 9700 AS, Groningen
The Netherlands
vdbeek@let.rug.nl

Abstract

This paper presents a range of methods for classifying Dutch noun countability based on either Dutch or English data. The classification is founded on translational equivalences and the corpus analysis of linguistic features which correlate with particular countability classes. We show that crosslingual classification on the basis of word-to-word or feature-to-feature mappings between English and Dutch performs at least as well as in-language classification based on gold-standard Dutch countability data.

1 Introduction

The performance of supervised learning methods is conditioned on the quality of annotation and also volume of training data (Hastie et al., 2001). This effect is felt particularly keenly in tasks of high feature dimensionality or low feature–class correlation. In many cases, high-quality data is not available in large quantities, but a large volume of lower-quality data can be accessed (Mitchell, 1999; Banko and Brill, 2001). Alternatively, high-quality data may exist for some parallel task which can be adapted to the task at hand through a lossy feature mapping. This strategy has been adopted successfully in NLP applications such as part-of-speech tagging involving languages with a relative paucity of language resources or annotated data (Yarowsky et al., 2001; Cucerzan and Yarowsky, 2002).

This paper takes a supervised learning task and contrasts the use of a restricted volume of in-language training data with the use of a larger volume of out-of-language training data adapted to the task through a lossy mapping. Our aim in this is to determine the most effective fast-track solution when faced with a novel task in a given language for which high-quality annotated data exists in a closely-related language.

We illustrate this issue by way of a type-level noun countability classification task in Dutch for which we have moderate amounts of high-quality annotated data in English and large amounts of medium-quality annotated data in Dutch (see §2.4). For English, previous research has shown that corpus evidence can be applied successfully to classify unannotated noun types for countability (Baldwin and Bond, 2003a; Baldwin and Bond, 2003b). We extend this research to Dutch and address the question of which of high-quality out-of-language English data and lower-quality in-language Dutch data produces the best Dutch countability classification results, realising that the feature mapping from English-to-Dutch in the first case will necessarily be lossy.

We treat **noun countability** as a lexical property that determines determiner co-occurrence, the ability to pluralise, and enumeration effects. Each Dutch noun type is classified as being countable and/or uncountable, noting that different senses/usages of a given word can occur with different countabilities, cf. *I want a rabbit* ⇌ *Ik will een konijn* [countable] vs. *I would like some more rabbit, please* ⇌ *I zou graag nog wat konijn willen* [uncountable]. Knowledge of countability is important both for analysis and generation. In analysis it helps to constrain the set of possible parses and their interpretation. In generation, countability information determines whether a noun can be pluralised and what determiners it can combine with.

The assumption underlying the crosslingual countability classification task is that Dutch and English are sufficiently close linguistically that there is a strong correlation between noun countability in the two languages. Both languages distinguish countable, uncountable and plural only nouns.[1] Although mismatches exist—e.g. *brain* [countable] vs. *hersenen* [plural only], *thun-*

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

derstorm [countable] vs. *onweer* [uncountable]—many Dutch words are in the same countability class as their English equivalents (e.g. *car* $\rightleftharpoons$ *auto* [countable], *food* $\rightleftharpoons$ *eten* [uncountable] and *goods* $\rightleftharpoons$ *goederen* [plural only]). 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, in the manner of Baldwin and Bond (2003a). 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 *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 is a weak indicator of plural only in English, but not in Dutch).

The remainder of this paper is structured as follows. §2 describes the countability classes, the nature and extraction of the features used in the corpus-based method, the feature abstraction method and the gold-standard data. §3 outlines the various classifiers tested in this research. §4 presents and discusses the experimental results. The conclusions of the paper are given in §5.

2 Preliminaries

2.1 Countability classes

Dutch and English nouns are generally considered to belong to one or more of three possible countability classes: countable, uncountable and plural only. **Countable** nouns can be modified by denumerators, prototypically numbers, and have a morphologically marked plural form: *one dog* $\rightleftharpoons$ *een hond*, *two dogs* $\rightleftharpoons$ *twee honden*. **Uncountable** nouns cannot be modified by denumerators, but can be modified by unspecific quantifiers such as *much* $\rightleftharpoons$ *veel*, and do not show any number distinction (prototypically being singular): **one rice* $\rightleftharpoons$ **een rijst*, *some rice* $\rightleftharpoons$ *een beetje rijst*, **two rices* $\rightleftharpoons$ **twee rijsten*. This class includes many abstract nouns, material-denoting nouns, generics and deverbalised nouns. **Plural only** nouns only have a plural form, such as *goods* $\rightleftharpoons$ *goederen* and cannot be denumerated. The plural only class is considered to be a closed class in Dutch, and is thus ignored in the classification experiments below.[2] Note that countability distinctions are in fact not categorical (Allan, 1980): prototypical countable nouns can be used in uncountable contexts, forcing a 'substance' interpretation (the **universal grinder**, e.g. *there was deer all over the road* $\rightleftharpoons$ *over de hele straat lag hert*) and uncountable nouns can in certain contexts be denumerated, resulting in a 'type' interpretation (the **universal packager**, e.g. *this shop sells three different wines* $\rightleftharpoons$ *deze winkel verkoopt drie verschillende wijnen*). However, nouns are generally considered to have a basic classification as countable and/or uncountable.

2.2 Feature space

The feature space used in this research 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 ($^{[x]_{\mathbf{L}}}$=1-dimensional feature cluster with x unit features for language **L**, $^{[x\times y]_{\mathbf{L}}}$= 2-dimensional feature cluster with $x \times y$ unit features for language **L**). 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: $^{[2]_{\mathbf{E}}} \rightleftharpoons {}^{[2]_{\mathbf{D}}}$ the number of the target noun when it heads an NP

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

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

N_1 *of* N_2/measure noun constructions: $^{[11\times2]_{\mathbf{E}}} \rightleftharpoons {}^{[11\times2]_{\mathbf{D}}}$ the type of the N_1 vs.

[2] But see van der Beek and Baldwin (2003) for classification results over the plural only class.

the number of the target noun (N_2) in an English N_1 *of* N_2 construction or Dutch measure noun construction. N_1 types include `COLLECTIVE` (*a group of people* $\rightleftharpoons$ *een groep mensen*), `UNIT` (*a kilo of sugar* $\rightleftharpoons$ *een kilo suiker*) and `TEMPORAL` (*a minute of silence* $\rightleftharpoons$ *een minuut stilte*).

Occurrence in PPs: $[52\times2]_{\mathbf{E}} \rightleftharpoons [84\times2]_{\mathbf{D}}$ the preposition type and presence or absence of a determiner when the target noun occurs in **singular** form in a PP.

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

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

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

Number-neutral determiners: $[11\times2]_{\mathbf{E}} \rightleftharpoons [13\times2]_{\mathbf{D}}$ what number-neutral determiners (e.g. *less*) 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, lemmata and chunk data were extracted from automatically generated, fully parsed **Alpino** output (Bouma et al., 2000). For English, we used a custom-built 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 Dutch and English in classification, we invariably combine like systems (e.g. Dutch POS data with English POS data).

The English data was extracted from the written component of the British National Corpus (90m words: Burnard (2000)), and the Dutch data from the newspaper component of the Twente Nieuws Corpus (20m words).[3]

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.

2.4 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. These two resources were combined in two ways: (1) by taking the intersection of positive and negative exemplars for each countability class (the **binary** datasets); and (2) by taking the union of all countabilities for a given word in the two resources and representing it as a single multiclass (i.e. countable, uncountable or countable+uncountable: the **multiclass** dataset). In each case, the total number of training instances is around 6,000 words. To determine the quality of annotation, we hand-annotated 100 unseen nouns and measured the agreement[4] with the gold-standard datasets. The agreement for the binary dataset was 92.4%, and that for the multiclass dataset was 89.8%.[5]

In Dutch, there are two electronic dictionaries with countability information: **CELEX** (Baayen et al., 1993) and the **Alpino** lexicon (Bouma et

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

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

[5] The disparity here is due to the fact that the binary dataset is constructed more conservatively and does not contain any words where there is disagreement in countability between **COMLEX** and **ALT-J/E**.

al., 2000). The latter includes the former as well as the Parole lexicon (no countability information), and has been manually modified and extended. We thus used the **Alpino** data to generate a total of three sets of training data for the monolingual Dutch classifiers in the same manner as for English: separate binary datasets for each of the countable and uncountable classes, and a combined multiclass-based dataset. The total number of training instances in each case is around 14,500, over twice the size of the English datasets.

In order to both evaluate the various classifiers and gauge the reliability of the **Alpino** countability judgements, we manually annotated 196 unseen Dutch nouns, basing judgements on actual usage in the Twente Nieuws Corpus. The agreement in countability judgements between the **Alpino** lexicon and hand-annotated data is 81.1%. This is markedly lower than the agreement for the English datasets, and supports our claims about the relatively low quality of the Dutch **Alpino** data as compared to the English data.

3 Classifier design

We propose a variety of both monolingual (Dutch-to–Dutch = **NN** and English-to-English = **EE**) and crosslingual (English-to-Dutch = **EN**) unsupervised and supervised classifier architectures for the task of learning countability. We employ two basic classifier architectures: (1) a separate binary classifier for each countability class (**BIN**), and (2) a single multiclass classifier (**MULTI**). In all cases, our supervised classifiers are built using TiMBL version 4.2 (Daelemans et al., 2002), a memory-based classification system based on the k-nearest neighbour algorithm.

3.1 Monolingual classifiers

Evidence-based classifiers: NN_{BIN}(evidence,$*$)

In an attempt to derive a baseline for each countability class/pre-processor system combination, we built a (binary) monolingual unsupervised classifier based on diagnostic evidence. For each target noun, the unsupervised classifier simply checks for the existence of diagnostic data in the output of each of the POS tagger and chunker for the given countability class (**NN(evidence,POS)** and **NN(evidence,chunk)**, respectively). Diagnostic data takes the form of unit features which are uniquely associated with a given countability class, e.g. the determiner *a* $\rightleftharpoons$ *een* co-occurring with a given (singular) noun is a strong indicator of that noun being countable. We perform basic system combination by positively classifying any noun for which either of the two pre-processors produces diagnostic data for the given countability class (**NN(evidence,all)**).

Distribution-based classifiers: NN_{BIN}(feat$_{\text{ALL}}$)

Despite our reservations about the quality of countability annotation in the **Alpino** lexicon, we implemented a conventional monolingual classifier based on the full feature set given above (§2.2). In this, 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 3 separate feature values:

$$corpfreq(f_s, w) = \frac{freq(f_s|w)}{freq(*)} \quad (1)$$

$$wordfreq(f_s, w) = \frac{freq(f_s|w)}{freq(w)} \quad (2)$$

$$featfreq(f_s, w) = \frac{freq(f_s|w)}{\sum_i freq(f_i|w)}\rangle \quad (3)$$

where $freq(*)$ is the frequency of all words in the corpus. That is, for each unit feature we capture the relative corpus frequency, frequency relative to the target word frequency, and frequency relative to other features in the same feature cluster. Thus, for an n-valued one-dimensional feature cluster, we generate $3n$ independent feature values.

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

In the case of a two-dimensional feature matrix (e.g. subject-position noun number vs. verb number agreement), each unit feature $f_{s,t}$ for target noun w is translated into $corpfreq(f_{s,t}, w)$, $wordfreq(f_{s,t}, w)$ and $featfreq(f_{s,t}, w)$ as above, and 2 additional feature values:

$$featdimfreq_1(f_{s,t}, w) = \frac{freq(f_{s,t}|w)}{\sum_i freq(f_{i,t}|w)} \quad (4)$$

$$featdimfreq_2(f_{s,t}, w) = \frac{freq(f_{s,t}|w)}{\sum_j freq(f_{s,j}|w)} \quad (5)$$

which represent the $featfreq$ values calculated along each of the two feature dimensions. As

for one-dimensional feature clusters, we introduce amalgamated features for each row ($f_{i,*}$) and column ($f_{*,j}$) of the feature matrix, and describe each in the form of 3 values. For further details, see the description of the monolingual English task in Baldwin and Bond (2003a). This abstraction generates a total of 1,664 individual feature values for Dutch.

We learned individual countable and uncountable classifiers from the binary **Alpino** data, averaging the feature values across those from the tagger and chunker in each case.[6]

3.2 Crosslingual classifiers

Translation-based classifier: $\mathbf{EN_{BIN}}$(translate)

Translation-based classification applies the observation that Dutch nouns often take the same countability as their English translation equivalents. First, we derive English countabilities from the binary gold-standard datasets supplemented with data from the output of a monolingual supervised English countability classifier ($\mathbf{EE_{BIN}(feat_{ALL})}$—see below). We then extract translation pairs from a bilingual dictionary (English–Dutch freedict version 1.1-1, containing 15,426 Dutch entries) and for each countability class, check for the existence of an English translation in the given countability class. If none of the English translations are classified as belonging to that countability class, we negatively classify the Dutch noun. In the event that no translation data exists for the Dutch noun or no countability data exists for the English translation(s), we classify the Dutch noun countability as unknown. Note that we map English plural only and bipartite nouns onto the Dutch uncountable class.

Transliteration-based classifier: $\mathbf{EN_{BIN}}$(transliterate)

Transliteration-based classification also applies the observation that countability is frequently preserved under translation from English to Dutch, but does so in a resource-free manner. It takes a Dutch noun and simply determines if a countability-annotated word of the same spelling exists in English, and if so, transfers the countability directly across to Dutch. In all other respects, we implement the method identically to translation-based classification.

[6] We additionally built separate classifiers based on the outputs of the individual pre-processors, and also based on the multiclass data, but found their performance to be marginally inferior to that of $\mathrm{NN_{BIN}}(\mathrm{feat_{ALL}})$.

Cluster-to-cluster classifier: $\mathbf{EN_{BIN}}$(cluster)

As observed above (§2.2), there is a strong correlation between the feature clusters used for Dutch and English. For example, co-occurrence with plural determiners is a strong indicator that the given noun is countable in both English and Dutch. At the same time, there is generally low correlation between individual unit features. For example, the English plural determiner *many* has no direct Dutch equivalent, and conversely, the Dutch plural determiner *sommige* has no direct English equivalent. The most straightforward way of aligning feature clusters, therefore, is through the (three) amalgamated totals for each one-dimensional feature cluster and some subset of the column and row 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 independent of number). All values for the individual unit features are then ignored. In this way, it is possible to align 88 feature values, based on the output of the English and Dutch POS taggers.[7] Note that as part of the feature alignment, we take the negative log of all corpus frequency ($corpfreq$) values in an attempt to reduce the effects of differing corpus sizes in English and Dutch.

Feature-to-feature classifiers: $\mathbf{EN_{*}(feat_{*})}$

While we stated above that there is generally low correlation between individual unit features in English and Dutch, some unit features are highly correlated crosslingually. One example is the English singular determiner *a* which correlates highly with the Dutch *een*. Here, we can thus simply match the feature values onto one another directly. In other cases, a many-to-many mapping exists between proper subsets of a given feature cluster (e.g. the English determiner pair *each* and *every* correlates highly with the Dutch determiner pair *ieder* and *elk*), and alignment takes the form of feature value amalgamation in each language by averaging over the unit values and aligning the amalgamated values. A total of 466 unit feature values are amalgamated into 351 feature values, which are then combined with the 88 aligned total values from cluster-to-cluster classification

[7] All crosslingual feature-based methods were tested over the output of the POS taggers, the chunkers and the combined outputs of the three English and two Dutch pre-processors. Overall, there was very little separating the results, and the simple POS tagger generally produced the most consistent results.

for a total of 439 feature values. As for cluster-to-cluster classification, we evaluate feature-to-feature classification over the output of the English and Dutch POS taggers.

We implemented a total of 5 feature-to-feature classifiers: (1) **EN$_{\text{BIN}}$(feat$_{\text{ALL}}$)** makes use of all aligned features in the form of separate binary classifiers; (2) **EN$_{\text{MULTI}}$(feat$_{\text{ALL}}$)** similarly uses all aligned features, but in a multiclass classifier architecture; (3) **EN$_{\text{BIN}}$(feat$_{\text{DET}}$)** is based on only aligned determiner features, plus the aligned cluster totals; (4) **EN$_{\text{BIN}}$(feat$_{\text{PREP}}$)** is based on only aligned preposition features, plus the aligned cluster totals; and (5) **EN$_{\text{BIN}}$(feat$_{\text{PRON}}$)** is based on only aligned pronoun features, plus the aligned cluster totals.[8]

3.3 System combination

System combination takes the outputs of heterogeneous classifiers and makes a consolidated classification based upon them. It has been shown to be effective in tasks ranging from word sense disambiguation to tagging in consolidating the performance of component systems (Klein et al., 2002; van Halteren et al., 2001). In our case, we first take the outputs of all unsupervised (i.e. evidence-based) and crosslingual classifiers—a total of 12 classifiers—for each countability class (**EN$_{\text{BIN}}$(combined)**). We test the effects of system classification by way of 10-fold cross-validation over the 196 annotated Dutch nouns. This provides an estimate of the classification performance we could expect over unannotated Dutch noun data using the 196 annotated nouns as our sole source of annotated Dutch data. We also test combining the outputs of the 12 unsupervised and crosslingual classifiers with that of the **Alpino**-trained Dutch classifier (**E/NN$_{\text{BIN}}$(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 baseline for each countability class is a majority-class binary classifier which simply classifies all instances according to the most commonly-attested class in the given dataset. Irrespective of the majority class, we calculate the recall and 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 (as occurs for uncountable nouns), the recall and F-score are given in parentheses.

We also provide an upper bound estimate of precision, recall and F-score based on a monolingual English countability classification task, with classifiers designed similarly to the monolingual Dutch classifiers (**EE$_{\text{BIN}}$(feat$_{\text{ALL}}$)**). In the case of English, the total number of feature values is 3,852, based on the concatenation of feature values from each of a POS tagger, chunker and dependency parser (Baldwin and Bond, 2003a). Our reason for choosing this as an upper bound is that it is based on moderate-volume, relatively noise-free training data and full feature correlation.

The classifier results are presented in Tables 1

Method	*Acc*	*P*	*R*	*F*
NN$_{\text{BIN}}$(majority)	.847	.847	1.000	.917
NN$_{\text{BIN}}$(evidence,all)	.551	.964	.488	.648
NN$_{\text{BIN}}$(evidence,chunk)	.510	.973	.434	.600
NN$_{\text{BIN}}$(evidence,POS)	.474	.970	.392	.558
EN$_{\text{BIN}}$(translate)	.948	.948	.331	.491
EN$_{\text{BIN}}$(transliterate)	**1.000**	**1.000**	.151	.262
EN$_{\text{BIN}}$(cluster)	.806	.957	.807	.876
EN$_{\text{MULTI}}$(cluster)	(.704)	.959	.837	.894
EN$_{\text{BIN}}$(feat$_{\text{ALL}}$)	.750	.983	.717	.829
EN$_{\text{MULTI}}$(feat$_{\text{ALL}}$)	(.704)	.959	.855	.904
EN$_{\text{BIN}}$(feat$_{\text{DET}}$)	.719	.966	.693	.807
EN$_{\text{BIN}}$(feat$_{\text{PREP}}$)	.755	.968	.735	.836
EN$_{\text{BIN}}$(feat$_{\text{PRON}}$)	.735	.952	.723	.822
EN$_{\text{BIN}}$(combined)	.873	.944	.904	.923
NN$_{\text{BIN}}$(feat$_{\text{ALL}}$)	.867	.961	.880	.918
E/NN$_{\text{BIN}}$(combined)	.903	.947	**.940**	**.943**
EE$_{\text{BIN}}$(feat$_{\text{ALL}}$)	—	.948	.972	.960

Table 1: Results for countable nouns

Method	*Acc*	*P*	*R*	*F*
NN$_{\text{BIN}}$(majority)	.638	.362	(1.000)	(.532)
NN$_{\text{BIN}}$(evidence,all)	.515	.423	**.930**	.581
NN$_{\text{BIN}}$(evidence,chunk)	.505	.414	.887	.565
NN$_{\text{BIN}}$(evidence,POS)	.628	.490	.718	.583
EN$_{\text{BIN}}$(translate)	.583	.583	.099	.169
EN$_{\text{BIN}}$(transliterate)	**.966**	**1.000**	.056	.107
EN$_{\text{BIN}}$(cluster)	.740	.692	.507	.585
EN$_{\text{MULTI}}$(cluster)	(.704)	.750	.592	.661
EN$_{\text{BIN}}$(feat$_{\text{ALL}}$)	.699	.750	.254	.379
EN$_{\text{MULTI}}$(feat$_{\text{ALL}}$)	(.704)	.822	.521	.638
EN$_{\text{BIN}}$(feat$_{\text{DET}}$)	.801	.758	.662	**.707**
EN$_{\text{BIN}}$(feat$_{\text{PREP}}$)	.776	.755	.563	.645
EN$_{\text{BIN}}$(feat$_{\text{PRON}}$)	.689	.708	.239	.358
EN$_{\text{BIN}}$(combined)	.791	.776	.593	.672
NN$_{\text{BIN}}$(feat$_{\text{ALL}}$)	.770	.783	.507	.615
E/NN(combined)	.812	.819	.609	.699
EE$_{\text{BIN}}$(feat$_{\text{ALL}}$)	—	.884	.907	.895

Table 2: Results for uncountable nouns

[8] Results for the multiclass classifier over feature subsets were found to be markedly worse than for binary classifiers.

and 2, broken down into the countable and uncountable classes. In each case, the best single value for each of evaluation metrics (other than the baseline and upper bound) is presented in **boldface**.

The first thing to notice is how much better the classifiers perform for countable than uncountable nouns. This is due to two factors: the relative occurrence of members of the two classes (as reflected in the majority class classification accuracies), and the relative volume of features correlated with each class. The relatively high baseline accuracy and F-score for countable nouns (.847 and .917) surpassed the performance of all classifiers other than the translation-based, transliteration-based, combined and monolingual classifiers. For uncountable nouns, on the other hand, appreciable gains over the baseline were observed for many of the systems. Results for countable nouns were relatively close to the upper bound results for the English monolingual classifier, whereas results for uncountable nouns were less competitive.

We have made the claim that, due to the lack of reliable training data in Dutch, crosslingual classification using English data is a viable option. This is borne out by the finding that the combined crosslingual classifier (EN_{BIN}(combined)) consistently outperforms the monolingual Dutch classifier in F-score, with the discrepancy being particularly noticeable for uncountable nouns. This finding is particularly striking given that the volume of Dutch training data is more than twice the volume of English data. Additional support comes from the analysis of the agreement between the system outputs the 196 hand-annotated nouns, recalling from §2.4 that the benchmark agreement for the **Alpino** data is 81.1%. The agreement for $NN_{BIN}(feat_{ALL})$ is 82.1%, that for EN_{BIN}(combined) is 83.2%, and that for E/NN_{BIN}(combined) is a respectable 85.7%. That is, all three methods produce countability judgements that are more parsimonious with actual corpus occurrence than the **Alpino** data, and the combined crosslingual classifier (EN_{BIN}(combined)) is superior to the monolingual classifier ($NN_{BIN}(feat_{ALL})$). Having said this, the combined crosslingual/monolingual classifier (E/NN_{BIN}(combined)) outperforms both the combined crosslingual classifier and the monolingual classifier, in which sense the **Alpino** data has some empirical utility. That is, we have shown that high-quality out-of-language English countability data is a stronger predictor of Dutch countability than medium-quality in-language Dutch countability data, but at the same time that the two are complementary.

There is very little separating the cluster-to-cluster and feature-to-feature classifiers. Given the high overhead in hand-aligning features in feature-to-feature classification, cluster-to-cluster classification would appear a low-cost, high-performance solution to the crosslingual countability task. Within the feature-to-feature classifiers, the results for the feature subsets are intriguing. We would expect that the determiner features should provide greater leverage than either the pronoun or preposition features, and this is indeed the case for uncountable nouns, where the determiner feature-based classifier returns the best F-score of all the classifiers. For countable nouns, however, the determiner features perform the worst of the three. Further research is required to determine the cause of this effect.

The results for the translation- and transliteration-based classifiers require qualification. Unlike the other classifiers, we do not get 100% coverage, as classification is possible only in the case that we have an English translation or transliteration with countability information. Strictly speaking, this diminished coverage should not be reflected in any of our evaluation metrics. In order to bring out this effect in Tables 1 and 2, we chose to base recall on the ratio of correctly-classified test exemplars to the number of positive-class exemplars, irrespective of whether the method is able to classify them. The F-score is thus proportionately low. If we were to base recall on the number of *classified* positive-class exemplars, the recall for the translation-based classifiers would become a perfect 1.000 ($\frac{55}{55}$) and 1.000 ($\frac{7}{7}$) for the countable and uncountable classes, respectively, and the corresponding numbers for the transliteration-based classifiers would be 1.000 ($\frac{25}{25}$) and 0.800 ($\frac{4}{5}$). That is, assuming we have English translation(s) for a Dutch noun or an English word of the same spelling, we get a very accurate estimate of the Dutch countability from the English countability data.

Finally, it is important to realise that these results are based on a limited test dataset (196 nouns) and that fuller evaluation is required to validate our findings. Also, our method relies crucially on the assumption that English and Dutch are closely-related languages, and its scalability

to alternate language pairs remains to be determined.

5 Conclusion

We have presented several methods for classifying Dutch nouns as countable and/or uncountable on the basis of Dutch and English data. The classifiers depend on translation/transliteration data or linguistic features that were extracted from unannotated corpora. We compared a range of crosslingual English-to-Dutch classifiers with a monolingual Dutch-to-Dutch classifier, and found that the crosslingual classifiers outperformed the monolingual classifier to varying degrees. Based on this, we suggest that the optimal fast-track solution to Dutch countability classification is to use English data.

In future research, we are interested in the possibility of co-training via translation- and transliteration-based classification, as this seems to provide a means for automatically generating high-quality Dutch countability data to learn a monolingual classifier from.

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 the anonymous reviewers for their valuable input on this research, and John Carroll for allowing us to use 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*, pages 463–70, Sapporo, Japan.

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)*, pages 73–80, Sapporo, Japan.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proc. of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*, Toulouse, France.

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

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.

Silviu Cucerzan and David Yarowsky. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan.

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`).

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.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *Elements of Statistical Learning*. Springer-Verlag, New York, USA.

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.

Tom M. Mitchell. 1999. The role of unlabeled data in supervised learning. In *Proc. of the Sixth International Colloquium on Cognitive Science*, San Sebastian, Spain.

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.

Leonoor van der Beek and Timothy Baldwin. 2003. Crosslingual countability classification: English meets Dutch. *LinGO Working Paper No. 2003-03*.

David Yarowsky, Grace Ngai, and R Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. of Human Language Technology (HLT) 2001*, pages 161–8, San Diego, USA.