

# Empirical Approaches to Multilingual Lexical Acquisition

Lecturer: Timothy Baldwin



THE UNIVERSITY OF  
MELBOURNE

# Lecture 6

## Crosslingual Countability Learning

# The Ins and Outs of Dutch Noun Countability Classification

(Baldwin and van der Beek 2003)

# Crosslinguistic Predictability of Countability

- In linguistically-related languages such as English and Dutch, countability generally patterns the same way:
  - ★ same basic behaviour of translation-equivalent lexical/syntactic markers of countability (e.g. *one dog*  $\Leftrightarrow$  *een hond*, *some rice*  $\Leftrightarrow$  *een beetje rijst*)
  - ★ translation pairs often have same countability:
    - \* *car*  $\Leftrightarrow$  *auto* [countable]
    - \* *food*  $\Leftrightarrow$  *eten* [uncountable]
    - \* BUT *thunderstorm* [countable] vs. *onweer* [uncountable]

# Mission Statement

- Investigate methods for learning Dutch noun **type** countability:
  - (a) **monolingually** based on Dutch corpus evidence
    - ★ direct port of Baldwin and Bond (2003) to Dutch
  - (b) **crosslingually** based on English countability data
    - ★ assumes that *similar* nouns in English and Dutch have the same countability

# Out-of- vs. In-language Classification

- Given high-quality training data in a closely-related language (English — **COMLEX** + **ALT-J/E**) and medium-quality data in the target language (Dutch — **Alpino** lexicon):
  - ★ which generates the best classifier?
  - ★ what is the best form of crosslingual mapping?
- Focus on the task of Dutch noun countability classification

## Underlying Point of Interest

- Known mismatches in English/Dutch countability, e.g. *thunderstorm* [countable] vs. *onweer* [uncountable]
  - English-to-Dutch countability mapping noisy
- What is the relative impact of the volume/quality of training data, plus noise in countability alignment?

	<b>Alpino</b>	<b>COMLEX+ALT-J/E</b>
<b>Volume</b>	High	Medium
<b>Quality</b>	Medium	High

# Approaches to Monolingual Classification

- **Evidence-based classification:** certain features are strong predictors of a unique countability class [baseline]
- **Corpus similarity:** words which occur with the same features in the same basic distribution tend to have the same countability
- **Distribution-based classification:** same as for EN-EN classification task (Baldwin and Bond (2003))

# Approaches to Crosslingual Classification

- **Corpus occurrence-based classification** (binary vs. multiclass):
  - ★ **cluster-to-cluster classification:** EN and ND feature clusters pattern the same
  - ★ **feature-to-feature classification:** EN and ND features pattern the same (all features vs. partitions of feature space)

- **Translation-based classification:** countability is preserved under translation (e.g. *car*  $\rightleftharpoons$  *auto* [countable])
- **Transliteration-based classification:** countability is preserved under transliteration (e.g. *paranoia*  $\rightleftharpoons$  *paranoia* [uncountable])
- **System combination:** classify according to combined outputs of individual methods
  - ★ crosslingual + unsupervised monolingual
  - ★ crosslingual + monolingual

# Evidence-based Classification (NN)

- **Method:** classify each noun according to **token** occurrence with any of a set of feature “triggers”, e.g.:

★  $a \Leftrightarrow \text{een} \rightarrow \text{countable}$

... *Cezanne snarling like a **dog** and then ...*

... *with a pack of **dogs** running beside them.*

★ bare singular  $\rightarrow \text{uncountable}$

*Amnesty International has received **information** ...*

*Recent **information** from former detainees ...*

# Corpus Similarity

- Identify lexical and/or constructional phenomena (feature clusters) which correlate with equivalent countability predictions for both English and Dutch
- Identify the unit features populating each feature cluster
- Determine the relative corpus occurrence of the features for each noun (when in NP head position)
- Inherit countability from training nouns with similar feature “signatures”

## Example Feature Clusters

**Coordinate noun number:**  $[2 \times 2]_E \Leftrightarrow [2 \times 2]_N$  target noun number vs. the number of the head nouns of conjuncts (e.g. dogs and mud =  $\langle \text{PLURAL}, \text{SINGULAR} \rangle$ )

**Occurrence in PPs:**  $[52 \times 2]_E \Leftrightarrow [84 \times 2]_N$  the presence or absence of a determiner ( $\pm \text{DET}$ ) when singular head complement of PP (e.g. per dog =  $\langle \text{per}, -\text{DET} \rangle$ ).

**Singular determiners:**  $[10]_E \Leftrightarrow [10]_N$  singular-selecting determiners (e.g. a dog = a)

# Example: Coordinate Noun Number

## Conjunct Number

		Conjunct Number	
		SINGULAR	PLURAL
Target Noun Number	SINGULAR	120	73
	PLURAL	2	5

$$\text{freq}(\ast) = 10^8$$

$$\text{freq}(n, \text{NP head}) = 10^4$$

# Distribution-based Classification (NN)

- Use full feature space (1,664 individual feature values)
- Train on **Alpino** data

# Cluster-to-cluster Distribution-based (EN)

- **Observation:** feature clusters in English and Dutch are basically equivalent in terms of their countability correlations, but individual features often do not match
- **Method:** align English and Dutch feature clusters via the cluster totals (88 feature values)
- Train on **COMPLEX+ALT-J/E** data

# Feature-to-feature Distribution-based (EN)

- **Observation:** some unit features or feature amalgams match closely between English and Dutch
- **Method:** distribution-based classification over a many-to-many mapping between English and Dutch features, resulting in 351 aligned feature values, e.g.:

★  $a \leftrightarrow een$

★  $\{each, every\} \leftrightarrow \{ieder, elk\}$

also use the 88 cluster-level feature values

## Example feature-to-feature classifiers

1. All aligned features, with  $2 \times$  binary classifiers
  2. Only aligned determiner features plus the aligned cluster totals, with  $2 \times$  binary classifiers
  - ⋮
- Total of 5 classifier configurations

# Translation-based Classification (EN)

- **Observation:** translation-equivalent nouns tend to have the same countability
- **Method:**
  - ★ Take union of the countabilities of translations for a given noun
  - ★ Countability **unknown** if no translations or no countability data for any of the translations
  - ★ Translations from English–Dutch freedict
  - ★ English countabilities from **COMLEX+ALT-J/E**, augmented with learned countabilities (31,111 nouns)

# Transliteration-based Classification (EN)

- **Observation:** there is significant lexical overlap between English and Dutch, and nouns of the same word form tend to have the same countabilities
- **Method:**
  - ★ Identical to translation-based classifier, except that countabilities based on the identical word form in English

# System Combination

- Test the following combinations of classifiers:
  - ★ all evidence-based and crosslingual classifiers (total of 12) for each countability class (**EN**)
  - ★ all monolingual and crosslingual classifiers (total of 13) for each countability class (**E/NN**)
- Combine by way of 10-fold cross-validation over the 196 annotated Dutch nouns

# Corpus Resources

- Source corpora:
  - ★ **English:** BNC (90m words)
  - ★ **Dutch:** Twente Nieuws Corpus (20m words)
- Pre-processors:
  - ★ POS tagger
  - ★ Full-text chunker
  - ★ Dependency parser (**English only**)

# Lexical Resources

- **Training data:**
  - ★ *English*: intersection of countabilities in **COMLEX** and **ALT-J/E** lexicons ( $\approx 6K$ )
  - ★ *Dutch*: countabilities in **Alpino** lexicon ( $\approx 14.5K$ )
- **Test data:** 196 Dutch nouns hand-annotated for countability based on corpus occurrence in Twente Nieuws Corpus (81.1% agreement with **Alpino**)

# Evaluation

- **Baseline:** majority-class classifier
- Evaluate according to classification accuracy, precision, recall and F-score
- Separate evaluation for countable and uncountable classes
- Compare against results for monolingual English countability classification task

# Results

- Better results for crosslingual than monolingual classification (!)
- Cluster-to-cluster classification  $\approx$  feature-to-feature classification
- Classifiers produce countability results more consistent with corpus occurrence than **Alpino** lexicon
- Translation and transliteration are excellent predictors of countability

# Ontology-based Crosslingual Countability Classification

(van der Beek and Baldwin 2004;  
van der Beek 2005)

# Ontology-based Classification

- Same basic approach as Bond and Vatikiotis-Bateson (2002), with the following extensions:
  - ★ use **EuroWordNet** as our ontology
  - ★ experiment with different combinations of languages (EN→EN, EN→ND, ND→EN, ND→ND)
  - ★ experiment with different options for deriving the countability prediction
  - ★ (optionally) use hypernyms and hyponyms in addition to synonyms

# Countability Classification Methods

- **Union-based classification:**

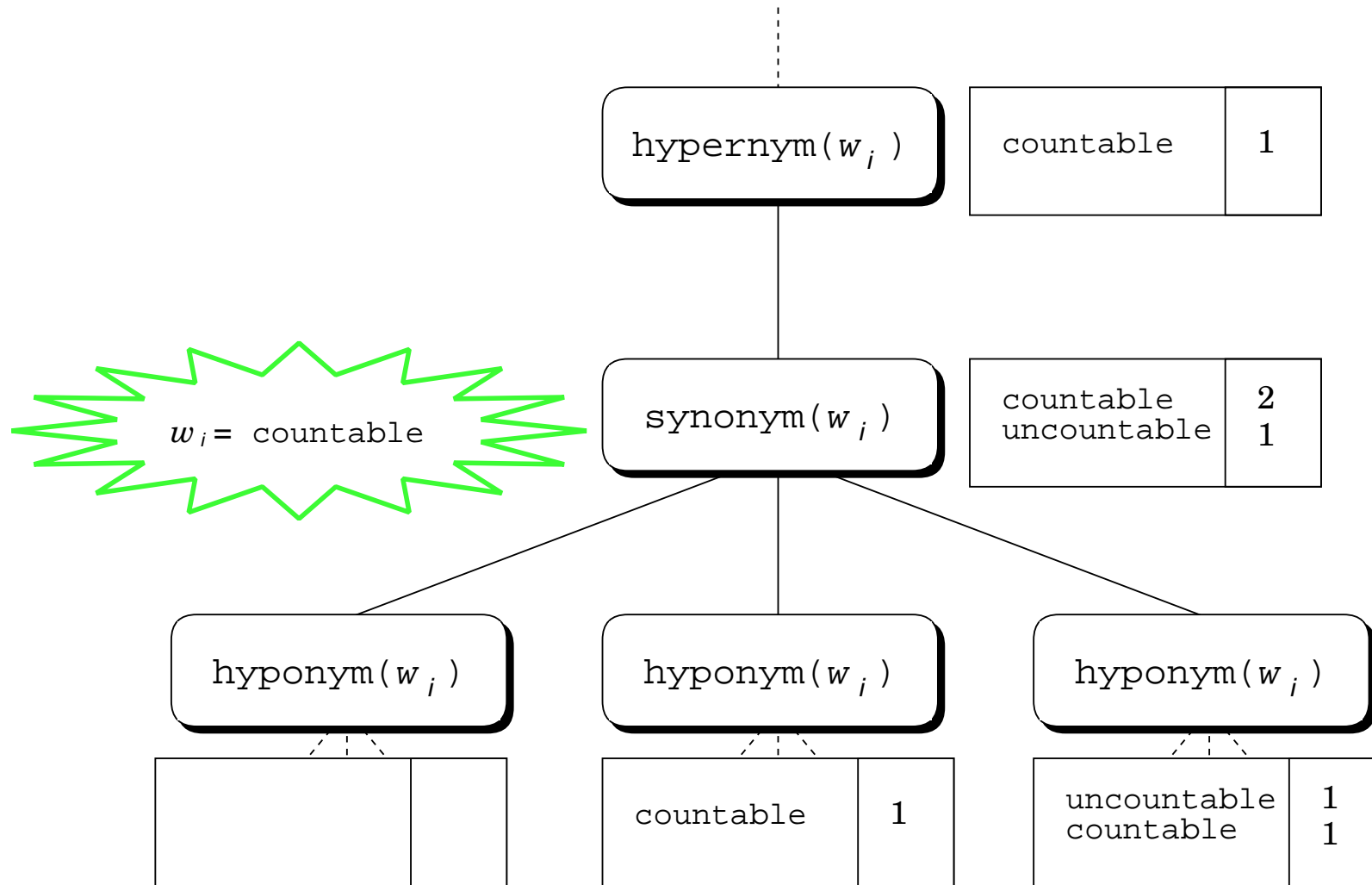
1. For each word sense  $w_i$  of word  $w$  in **EuroWordNet**:
  - ★ take **union** of all countabilities of near-neighbour training words
2. Take union of lexical types for all  $w_i$

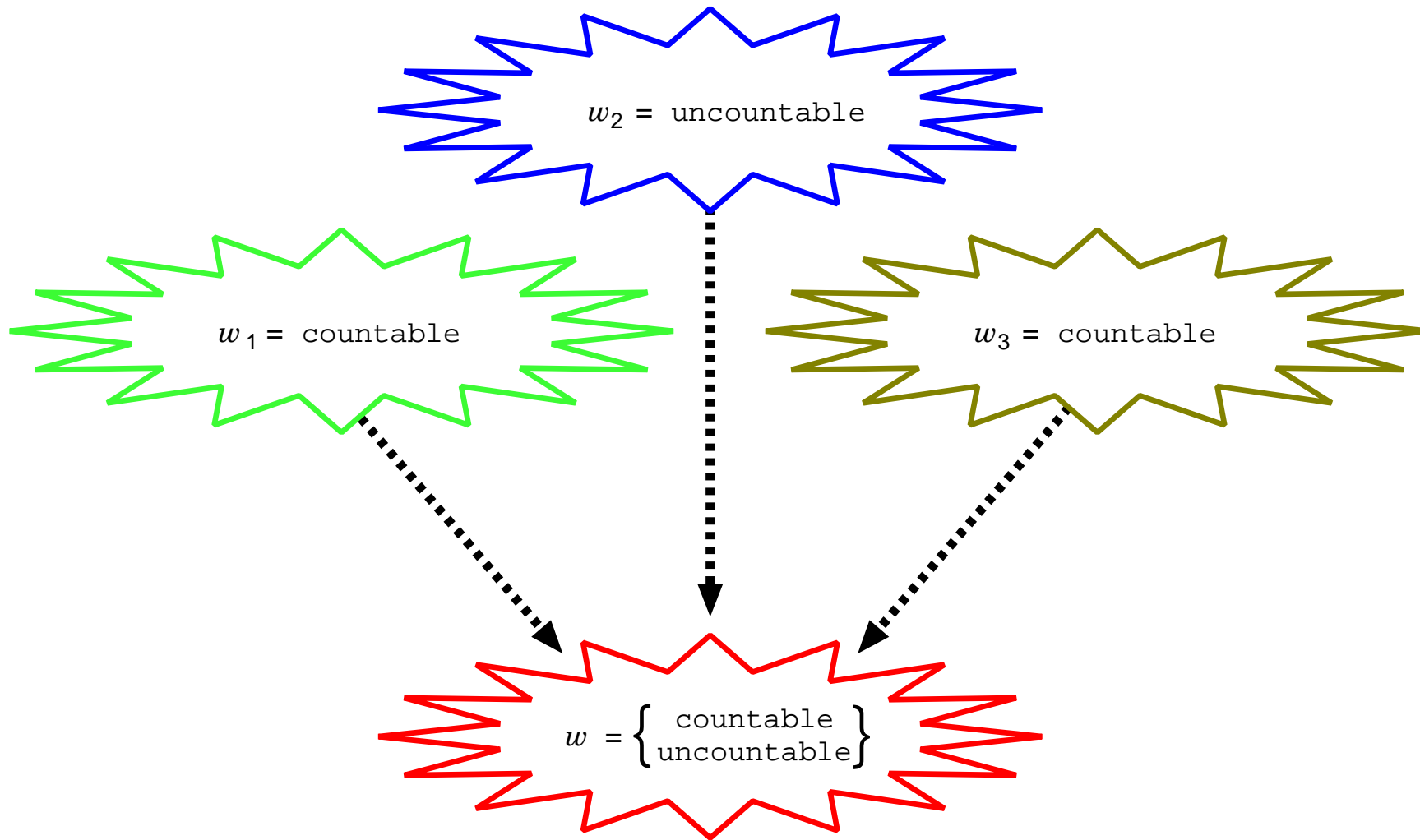
- **Majority-based classification**

1. For each word sense  $w_i$  of word  $w$  in **EuroWordNet**:
  - ★ count up the number of votes for each class based on the countabilities of near-neighbour training words
2. Return the (unique) majority vote =  $\arg \max_c \sum_i freq_c(w_i)$

- **Combined classification:**

1. For each word sense  $w_i$  of word  $w$  in **EuroWordNet**:
  - ★ count up the number of votes for each class based on the countabilities of near-neighbour training words, and determine the majority vote for  $w_i$  based on  $\arg \max_c \text{freq}_c(w_i)$
2. Take the union of classes across all  $w_i$





# Ontological Links

- **Synonymy-based classification:** use only synonyms (a lá Bond and Vatikiotis-Bateson (2002))
- **Hypernym-based classification:** base classification on synonyms, with hypernym backoff
- **Hyponym-based classification:** base classification on synonyms, with hyponym backoff
- **Bidirectional classification:** base classification on synonyms, with hypernym+hyponym backoff

# Results

- Combined classification  $>$  majority-based  $\gg$  union-based
- Bidirectional  $>$  hypernym-based  $>$  hyponym-based  $\gg$  synonym-based (in terms of F-score)
- Better results for EN $\rightarrow$ ND than ND $\rightarrow$ ND, and ND $\rightarrow$ EN than EN $\rightarrow$ EN (!!)
- Best results for EN/ND $\rightarrow$ ND and EN/ND $\rightarrow$ EN

# Reflections

- Demonstration of types of methods that can be used to determine noun type countability
  - ★ distribution-based
  - ★ semantics/sense-based
  - ★ translation/transliteration-based

## Overall Caveats

- Results conditional on there being a close linguistic correspondence between the languages in question
- Highly selective use made of English countability data as compared to Dutch data
- Dutch countability data shown to boost the performance of the combined classifier

# References

- BALDWIN, TIMOTHY, and LEONOR VAN DER BEEK. 2003. The ins and outs of Dutch noun countability classification. In *Proc. of the 2003 Australasian Language Technology Workshop (ALTW2003)*, 33–40, Melbourne, Australia.
- BOND, FRANCIS, and CAITLIN VATIKIOTIS-BATESON. 2002. Using an ontology to determine English countability. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.
- VAN DER BEEK, LEONOR, 2005. *Topics in Corpus-Based Dutch Syntax*. University of Groningen dissertation.
- , and TIMOTHY BALDWIN. 2004. Crosslingual countability classification with EuroWordNet. In *Papers from the 14th Meeting of Computational Linguistics in the Netherlands*, 141–55, Antwerp, Belgium. Antwerp Papers in Linguistics.