# Computational Approaches to Translation Studies

Shuly Wintner

Department of Computer Science
University of Haifa
Haifa, Israel
shuly@cs.haifa.ac.il

# Original or Translation?

## Example (O or T?)

*We want to see countries that can produce the best product for the best price in that particular business . I have to agree with the member that free trade agreements by definition do not mean that we have to be less vigilant all of a sudden .*

## Example (T or O?)

*I would like as my final point to say that we support free trade , but we must learn from past mistakes . Let us hope that negotiations for free trade agreements with the four Central American countries introduce a number of other dimensions absent from these first generation agreements .*

# Underlying Assumptions

- Research in Translation Studies can inform Natural Language Processing, and in particular improve the quality of machine translation
- Computational methodologies can shed light on pertinent questions in Translation Studies

# Outline

# Translationese

## The language of translated texts

- Translated texts differ from original ones
- The differences do not indicate poor translation but rather a statistical phenomenon, **translationese** (Gellerstam, 1986)
- Toury (1980, 1995) defines two **laws of translation**:
  - The Law of Interference: Fingerprints of the source text that are left in the translation product
  - The Law of Standardization: Adjust the translation product according to existing norms in the target language and culture

# Translationese

## The language of translated texts

- Translated texts differ from original ones
- The differences do not indicate poor translation but rather a statistical phenomenon, **translationese** (Gellerstam, 1986)
- Toury (1980, 1995) defines two **laws of translation**:

  The law of interference Fingerprints of the source text that are left in the translation product

  The law of growing standardization Effort to standardize the translation product according to existing norms in the target language and culture

# Translationese

## The language of translated texts

- Translated texts differ from original ones
- The differences do not indicate poor translation but rather a statistical phenomenon, **translationese** (Gellerstam, 1986)
- Toury (1980, 1995) defines two **laws of translation**:

  The law of interference Fingerprints of the source text that are left in the translation product

  The law of growing standardization Effort to standardize the translation product according to existing norms in the target language and culture

# Translationese

## The language of translated texts

- Translated texts differ from original ones
- The differences do not indicate poor translation but rather a statistical phenomenon, **translationese** (Gellerstam, 1986)
- Toury (1980, 1995) defines two **laws of translation**:

  The law of interference Fingerprints of the source text that are left in the translation product

  The law of growing standardization Effort to standardize the translation product according to existing norms in the target language and culture

# Translationese

## The language of translated texts

- Translated texts differ from original ones
- The differences do not indicate poor translation but rather a statistical phenomenon, **translationese** (Gellerstam, 1986)
- Toury (1980, 1995) defines two **laws of translation**:

  The law of interference Fingerprints of the source text that are left in the translation product

  The law of growing standardization Effort to standardize the translation product according to existing norms in the target language and culture

# Translationese

## The language of translated texts

### Translation universals (Baker, 1993)

> *"features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems"*

- simplification (Blum-Kulka and Levenston, 1978, 1983)
- explicitation (Blum-Kulka, 1986)
- normalization (Chesterman, 2004)

# Translationese

## The language of translated texts

**Translation universals** (Baker, 1993)

> *"features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems"*

**simplification** (Blum-Kulka and Levenston, 1978, 1983)

**explicitation** (Blum-Kulka, 1986)

**Normalization** (Chesterman, 2004)

# Translationese

## The language of translated texts

**Translation universals** (Baker, 1993)

> *"features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems"*

**simplification** (Blum-Kulka and Levenston, 1978, 1983)

**explicitation** (Blum-Kulka, 1986)

**Normalization** (Chesterman, 2004)

# Translationese

## The language of translated texts

Translation universals (Baker, 1993)

> *"features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems"*

simplification (Blum-Kulka and Levenston, 1978, 1983)

explicitation (Blum-Kulka, 1986)

Normalization (Chesterman, 2004)

# Translationese

## The language of translated texts

Translation universals (Baker, 1993)

> *"features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems"*

simplification (Blum-Kulka and Levenston, 1978, 1983)

explicitation (Blum-Kulka, 1986)

Normalization (Chesterman, 2004)

# Translationese

## Why does it matter?

- Language models for statistical machine translation (Lembersky et al., 2011, 2012b)
- Translation models for statistical machine translation (Kurokawa et al., 2009; Lembersky et al., 2012a, 2013)
- Cleaning parallel corpora crawled from the Web (Eetemadi and Toutanova, 2014; Aharoni et al., 2014)
- Understanding the properties of human translation (Ilisei et al., 2010; Ilisei and Inkpen, 2011; Ilisei, 2013; Volansky et al., 2015; Avner et al., 2016)

# Translationese

## Why does it matter?

- Language models for statistical machine translation (Lembersky et al., 2011, 2012b)
- Translation models for statistical machine translation (Kurokawa et al., 2009; Lembersky et al., 2012a, 2013)
- Cleaning parallel corpora crawled from the Web (Eetemadi and Toutanova, 2014; Aharoni et al., 2014)
- Understanding the properties of human translation (Ilisei et al., 2010; Ilisei and Inkpen, 2011; Ilisei, 2013; Volansky et al., 2015; Avner et al., 2016)

# Translationese

## Why does it matter?

- Language models for statistical machine translation (Lembersky et al., 2011, 2012b)
- Translation models for statistical machine translation (Kurokawa et al., 2009; Lembersky et al., 2012a, 2013)
- Cleaning parallel corpora crawled from the Web (Eetemadi and Toutanova, 2014; Aharoni et al., 2014)
- Understanding the properties of human translation (Ilisei et al., 2010; Ilisei and Inkpen, 2011; Ilisei, 2013; Volansky et al., 2015; Avner et al., 2016)

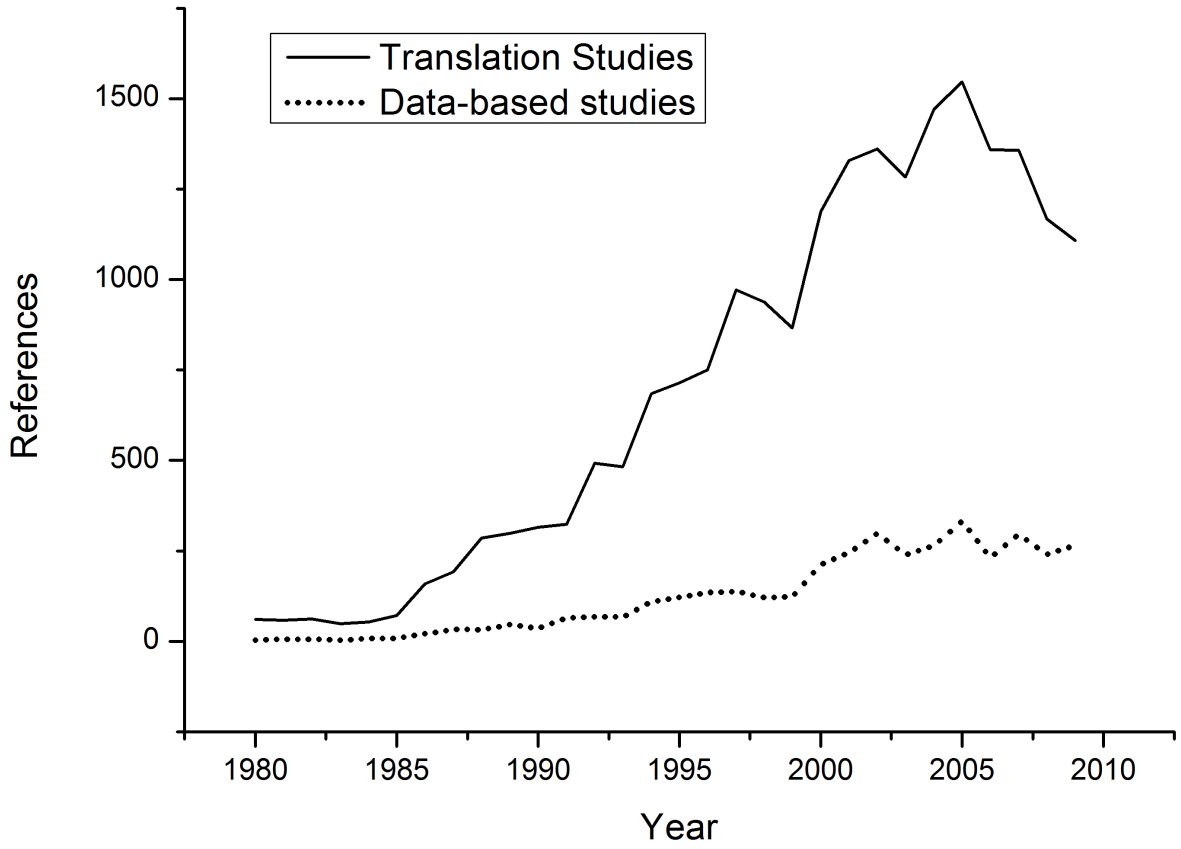
# Translationese

## Why does it matter?

- Language models for statistical machine translation (Lembersky et al., 2011, 2012b)
- Translation models for statistical machine translation (Kurokawa et al., 2009; Lembersky et al., 2012a, 2013)
- Cleaning parallel corpora crawled from the Web (Eetemadi and Toutanova, 2014; Aharoni et al., 2014)
- Understanding the properties of human translation (Ilisei et al., 2010; Ilisei and Inkpen, 2011; Ilisei, 2013; Volansky et al., 2015; Avner et al., 2016)

# Corpus-based Translation Studies

## Example (Sun and Shreve (2013))



# Computational Investigation of Translationese

- Translated texts exhibit lower lexical variety (type-to-token ratio) than originals (Al-Shabab, 1996)
- Their mean sentence length and lexical density (ratio of content to non-content words) are lower (Laviosa, 1998)
- Corpus-based evidence for the simplification hypothesis (Laviosa, 2002)

# Computational Investigation of Translationese

- Translated texts exhibit lower lexical variety (type-to-token ratio) than originals (Al-Shabab, 1996)
- Their mean sentence length and lexical density (ratio of content to non-content words) are lower (Laviosa, 1998)
- Corpus-based evidence for the simplification hypothesis (Laviosa, 2002)

# Computational Investigation of Translationese

- Translated texts exhibit lower lexical variety (type-to-token ratio) than originals (Al-Shabab, 1996)
- Their mean sentence length and lexical density (ratio of content to non-content words) are lower (Laviosa, 1998)
- Corpus-based evidence for the simplification hypothesis (Laviosa, 2002)

# Computational Investigation of Translationese

## Example (Punctuation marks across O and T)

| Mark | Frequency O | Frequency T | Ratio | LL | Weight |
|---|---|---|---|---|---|
| , | 37.83 | 49.79 | 0.76 | T | T1 |
| ( | 0.42 | 0.72 | 0.58 | T | T2 |
| ' | 1.94 | 2.53 | 0.77 | T | T3 |
| ) | 0.40 | 0.72 | 0.56 | T | T4 |
| / | 0.30 | 0.30 | 1.00 | — | — |
| [ | 0.01 | 0.02 | 0.45 | T | — |
| ] | 0.01 | 0.02 | 0.44 | T | — |
| " | 0.33 | 0.22 | 1.46 | O | O7 |
| ! | 0.22 | 0.17 | 1.25 | O | O6 |
| . | 38.20 | 34.60 | 1.10 | O | O5 |
| : | 1.20 | 1.17 | 1.03 | — | O4 |
| ; | 0.84 | 0.83 | 1.01 | — | O3 |
| ? | 1.57 | 1.11 | 1.41 | O | O2 |
| - | 2.68 | 2.25 | 1.19 | O | O1 |

# Outline

# Methodology

- Corpus-based approach
- Text classification with (supervised) machine-learning techniques
- Feature design
- Evaluation: ten-fold cross-validation
- Unsupervised classification

# Methodology

- Corpus-based approach
- Text classification with (supervised) machine-learning techniques
- Feature design
- Evaluation: ten-fold cross-validation
- Unsupervised classification

# Methodology

- Corpus-based approach
- Text classification with (supervised) machine-learning techniques
- Feature design
- Evaluation: ten-fold cross-validation
- Unsupervised classification

# Methodology

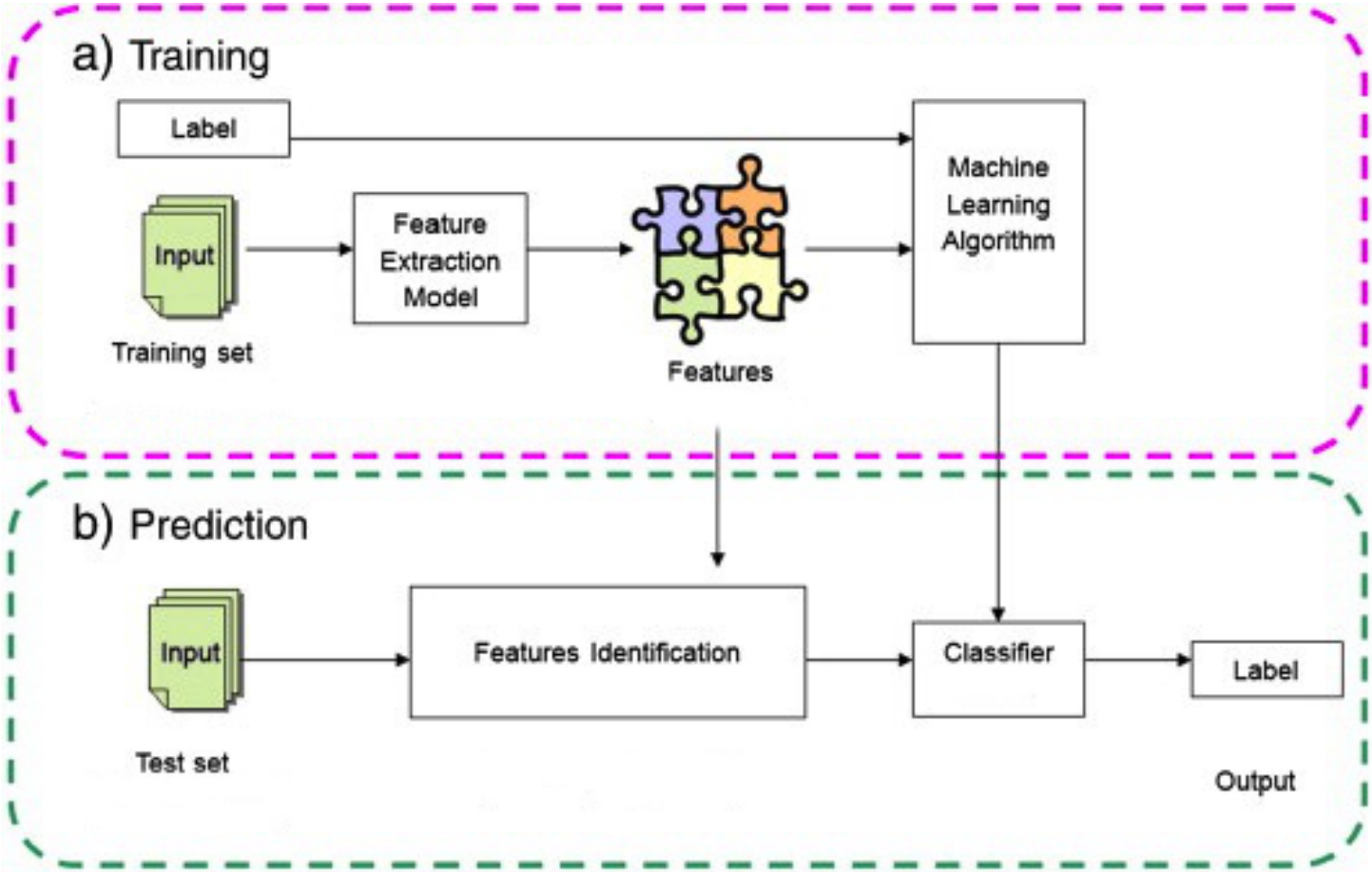
- Corpus-based approach
- Text classification with (supervised) machine-learning techniques
- Feature design
- Evaluation: ten-fold cross-validation
- Unsupervised classification

# Methodology

- Corpus-based approach
- Text classification with (supervised) machine-learning techniques
- Feature design
- Evaluation: ten-fold cross-validation
- **Un**supervised classification

# Text Classification with Machine Learning

## Example (From He et al. (2012))



# Text Classification with Machine Learning

- **Supervised machine learning**: a **training** corpus lists **instances** of both classes
- Each instance in the two classes is **represented** by a set of numeric **features** that are extracted from the instances
- A generic machine-learning algorithm is trained to distinguish between **feature vectors** representative of one class and those representative of the other
- The trained classifier is tested on an 'unseen' text, the **test set**
- Classifiers assign **weights** to the features

# Text Classification (Authorship Attribution)

## Applications

- Determine the gender/age of the author (Koppel et al., 2003)
- Tell Shakespeare from Marlowe (Juola, 2006)
- Identify suicide letters
- Spot plagiarism
- Filter out spam
- Identify the native language of non-native authors (Tetreault et al., 2013)

# Identifying Translationese

## Using text classification

- Baroni and Bernardini (2006)
- van Halteren (2008)
- Kurokawa et al. (2009)
- Ilisei et al. (2010); Ilisei and Inkpen (2011); Ilisei (2013)
- Koppel and Ordan (2011)
- Popescu (2011)
- Volansky et al. (2015); Avner et al. (2016)

# Experimental Setup

- Corpus: EUROPARL (Koehn, 2005)
- 4 million tokens in English (O)
- 400,000 tokens translated from each of the ten source languages (T): Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish
- The corpus is tokenized and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary)
- Part-of-speech tagging
- Classification with Weka (Hall et al., 2009), using SVM with a default linear kernel

# Experimental Setup

- Corpus: EUROPARL (Koehn, 2005)
- 4 million tokens in English (O)
- 400,000 tokens translated from each of the ten source languages (T): Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish
- The corpus is tokenized and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary)
- Part-of-speech tagging
- Classification with Weka (Hall et al., 2009), using SVM with a default linear kernel

# Experimental Setup

- Corpus: EUROPARL (Koehn, 2005)
- 4 million tokens in English (O)
- 400,000 tokens translated from each of the ten source languages (T): Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish
- The corpus is tokenized and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary)
- Part-of-speech tagging
- Classification with Weka (Hall et al., 2009), using SVM with a default linear kernel

# Experimental Setup

- Corpus: EUROPARL (Koehn, 2005)
- 4 million tokens in English (O)
- 400,000 tokens translated from each of the ten source languages (T): Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish
- The corpus is tokenized and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary)
- Part-of-speech tagging
- Classification with Weka (Hall et al., 2009), using SVM with a default linear kernel

# Experimental Setup

- Corpus: EUROPARL (Koehn, 2005)
- 4 million tokens in English (O)
- 400,000 tokens translated from each of the ten source languages (T): Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish
- The corpus is tokenized and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary)
- Part-of-speech tagging
- Classification with Weka (Hall et al., 2009), using SVM with a default linear kernel

# Experimental Setup

- Corpus: EUROPARL (Koehn, 2005)
- 4 million tokens in English (O)
- 400,000 tokens translated from each of the ten source languages (T): Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish
- The corpus is tokenized and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary)
- Part-of-speech tagging
- Classification with Weka (Hall et al., 2009), using SVM with a default linear kernel

# Outline

# Hypotheses

**Simplification** Rendering complex linguistic features in the source text into simpler features in the target (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985; Baker, 1993)

**Explicitation** The tendency to spell out in the target text utterances that are more implicit in the source (Blum-Kulka, 1986; Øverås, 1998; Baker, 1993)

**Normalization** Efforts to standardize texts (Toury, 1995), "a strong preference for **conventional grammaticality**" (Baker, 1993)

**Interference** The fingerprints of the source language on the translation output (Toury, 1979)

**Miscellaneous**

# Hypotheses

**Simplification** Rendering complex linguistic features in the source text into simpler features in the target (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985; Baker, 1993)

**Explicitation** The tendency to spell out in the target text utterances that are more implicit in the source (Blum-Kulka, 1986; Øverås, 1998; Baker, 1993)

**Normalization** Efforts to standardize texts (Toury, 1995), "a strong preference for **conventional grammaticality**" (Baker, 1993)

**Interference** The fingerprints of the source language on the translation output (Toury, 1979)

**Miscellaneous**

# Hypotheses

**Simplification** Rendering complex linguistic features in the source text into simpler features in the target (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985; Baker, 1993)

**Explicitation** The tendency to spell out in the target text utterances that are more implicit in the source (Blum-Kulka, 1986; Øverås, 1998; Baker, 1993)

**Normalization** Efforts to standardize texts (Toury, 1995), "a strong preference for **conventional grammaticality**" (Baker, 1993)

**Interference** The fingerprints of the source language on the translation output (Toury, 1979)

**Miscellaneous**

# Hypotheses

**Simplification** Rendering complex linguistic features in the source text into simpler features in the target (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985; Baker, 1993)

**Explicitation** The tendency to spell out in the target text utterances that are more implicit in the source (Blum-Kulka, 1986; Øverås, 1998; Baker, 1993)

**Normalization** Efforts to standardize texts (Toury, 1995), "a strong preference for **conventional grammaticality**" (Baker, 1993)

**Interference** The fingerprints of the source language on the translation output (Toury, 1979)

**Miscellaneous**

# Hypotheses

**Simplification** Rendering complex linguistic features in the source text into simpler features in the target (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985; Baker, 1993)

**Explicitation** The tendency to spell out in the target text utterances that are more implicit in the source (Blum-Kulka, 1986; Øverås, 1998; Baker, 1993)

**Normalization** Efforts to standardize texts (Toury, 1995), "a strong preference for **conventional grammaticality**" (Baker, 1993)

**Interference** The fingerprints of the source language on the translation output (Toury, 1979)

**Miscellaneous**

# Features Should...

1. Reflect frequent linguistic characteristics we would expect to be present in the two types of text
2. Be content-independent, indicating formal and stylistic differences between the texts that are not derived from differences in contents, domain, genre, etc.
3. Be easy to interpret, yielding insights regarding the differences between original and translated texts

# Features Should...

1. Reflect frequent linguistic characteristics we would expect to be present in the two types of text
2. Be content-independent, indicating formal and stylistic differences between the texts that are not derived from differences in contents, domain, genre, etc.
3. Be easy to interpret, yielding insights regarding the differences between original and translated texts

# Features Should...

1. Reflect frequent linguistic characteristics we would expect to be present in the two types of text
2. Be content-independent, indicating formal and stylistic differences between the texts that are not derived from differences in contents, domain, genre, etc.
3. Be easy to interpret, yielding insights regarding the differences between original and translated texts

# Simplification

**Lexical Variety** Three different **type-token ratio** (TTR) measures, where $V$ is the number of types and $N$ is the number of tokens per chunk, $V_1$ is the number of types occurring only once in the chunk

$$V/N \qquad \log(V)/\log(N) \qquad 100 \times \log(N)/(1 - v_1/v)$$

**Mean word length** In characters

**Syllable ratio** the number of vowel-sequences that are delimited by consonants or space in a word

**Mean sentence length** In words

**Lexical density** The frequency of tokens that are **not** nouns, adjectives, adverbs or verbs

**Mean word rank** The average **rank** of words in a frequency-ordered list of 6000 English most frequent words

**Most frequent words** The frequencies of the $N$ most frequent words, $N = 5, 10, 50$

# Simplification

**Lexical Variety** Three different **type-token ratio** (TTR) measures, where $V$ is the number of types and $N$ is the number of tokens per chunk, $V_1$ is the number of types occurring only once in the chunk

$$V/N \qquad \log(V)/\log(N) \qquad 100 \times \log(N)/(1 - V_1/V)$$

**Mean word length** In characters

**Syllable ratio** the number of vowel-sequences that are delimited by consonants or space in a word

**Mean sentence length** In words

**Lexical density** The frequency of tokens that are **not** nouns, adjectives, adverbs or verbs

**Mean word rank** The average **rank** of words in a frequency-ordered list of 6000 English most frequent words

**Most frequent words** The frequencies of the $N$ most frequent words, $N = 5, 10, 50$

# Simplification

**Lexical Variety** Three different **type-token ratio** (TTR) measures, where $V$ is the number of types and $N$ is the number of tokens per chunk, $V_1$ is the number of types occurring only once in the chunk

$$V/N \qquad \log(V)/\log(N) \qquad 100 \times \log(N)/(1 - V_1/V)$$

**Mean word length** In characters

**Syllable ratio** the number of vowel-sequences that are delimited by consonants or space in a word

**Mean sentence length** In words

**Lexical density** The frequency of tokens that are **not** nouns, adjectives, adverbs or verbs

**Mean word rank** The average **rank** of words in a frequency-ordered list of 6000 English most frequent words

**Most frequent words** The frequencies of the $N$ most frequent words, $N = 5, 10, 50$

# Simplification

**Lexical Variety** Three different **type-token ratio** (TTR) measures, where $V$ is the number of types and $N$ is the number of tokens per chunk, $V_1$ is the number of types occurring only once in the chunk

$$V/N \qquad \log(V)/\log(N) \qquad 100 \times \log(N)/(1 - V_1/V)$$

**Mean word length** In characters

**Syllable ratio** the number of vowel-sequences that are delimited by consonants or space in a word

**Mean sentence length** In words

**Lexical density** The frequency of tokens that are **not** nouns, adjectives, adverbs or verbs

**Mean word rank** The average **rank** of words in a frequency-ordered list of 6000 English most frequent words

**Most frequent words** The frequencies of the $N$ most frequent words, $N = 5, 10, 50$

# Simplification

**Lexical Variety** Three different **type-token ratio** (TTR) measures, where $V$ is the number of types and $N$ is the number of tokens per chunk, $V_1$ is the number of types occurring only once in the chunk

$$V/N \qquad \log(V)/\log(N) \qquad 100 \times \log(N)/(1 - V_1/V)$$

**Mean word length** In characters

**Syllable ratio** the number of vowel-sequences that are delimited by consonants or space in a word

**Mean sentence length** In words

**Lexical density** The frequency of tokens that are **not** nouns, adjectives, adverbs or verbs

**Mean word rank** The average **rank** of words in a frequency-ordered list of 6000 English most frequent words

**Most frequent words** The frequencies of the $N$ most frequent words, $N = 5, 10, 50$

# Simplification

**Lexical Variety** Three different **type-token ratio** (TTR) measures, where $V$ is the number of types and $N$ is the number of tokens per chunk, $V_1$ is the number of types occurring only once in the chunk

$$V/N \qquad \log(V)/\log(N) \qquad 100 \times \log(N)/(1 - V_1/V)$$

**Mean word length** In characters

**Syllable ratio** the number of vowel-sequences that are delimited by consonants or space in a word

**Mean sentence length** In words

**Lexical density** The frequency of tokens that are **not** nouns, adjectives, adverbs or verbs

**Mean word rank** The average **rank** of words in a frequency-ordered list of 6000 English most frequent words

**Most frequent words** The frequencies of the $N$ most frequent words, $N = 5, 10, 50$

# Simplification

**Lexical Variety** Three different **type-token ratio** (TTR) measures, where $V$ is the number of types and $N$ is the number of tokens per chunk, $V_1$ is the number of types occurring only once in the chunk

$$V/N \qquad \log(V)/\log(N) \qquad 100 \times \log(N)/(1 - V_1/V)$$

**Mean word length** In characters

**Syllable ratio** the number of vowel-sequences that are delimited by consonants or space in a word

**Mean sentence length** In words

**Lexical density** The frequency of tokens that are **not** nouns, adjectives, adverbs or verbs

**Mean word rank** The average **rank** of words in a frequency-ordered list of 6000 English most frequent words

**Most frequent words** The frequencies of the $N$ most frequent words, $N = 5, 10, 50$

# Explicitation

## Example (Explicitation)

| T | O |
|---|---|
| *israelische Ministerpräsident Benjamin Netanjahu* | *Merkel* |

**Explicit naming** The ratio of personal pronouns to proper nouns; this models the tendency to spell out pronouns

**Single naming** The frequency of proper nouns consisting of a single token, not having an additional proper noun as a neighbor

**Mean multiple naming** The average length (in tokens) of proper nouns

**Cohesive markers** The frequencies of several **cohesive markers** (*because*, *but*, *hence*, *in fact*, *therefore*, ...)

# Explicitation

## Example (Explicitation)

| T | O |
|---|---|
| *israelische Ministerpräsident Benjamin Netanjahu* | *Merkel* |

**Explicit naming** The ratio of personal pronouns to proper nouns; this models the tendency to spell out pronouns

**Single naming** The frequency of proper nouns consisting of a single token, not having an additional proper noun as a neighbor

**Mean multiple naming** The average length (in tokens) of proper nouns

**Cohesive markers** The frequencies of several **cohesive markers** (*because*, *but*, *hence*, *in fact*, *therefore*, ...)

# Explicitation

## Example (Explicitation)

| T | O |
|---|---|
| *israelische Ministerpräsident Benjamin Netanjahu* | *Merkel* |

**Explicit naming** The ratio of personal pronouns to proper nouns; this models the tendency to spell out pronouns

**Single naming** The frequency of proper nouns consisting of a single token, not having an additional proper noun as a neighbor

**Mean multiple naming** The average length (in tokens) of proper nouns

**Cohesive markers** The frequencies of several **cohesive markers** (*because*, *but*, *hence*, *in fact*, *therefore*, ...)

# Explicitation

## Example (Explicitation)

| T | O |
|---|---|
| *israelische Ministerpräsident Benjamin Netanjahu* | *Merkel* |

**Explicit naming** The ratio of personal pronouns to proper nouns; this models the tendency to spell out pronouns

**Single naming** The frequency of proper nouns consisting of a single token, not having an additional proper noun as a neighbor

**Mean multiple naming** The average length (in tokens) of proper nouns

**Cohesive markers** The frequencies of several **cohesive markers** (*because*, *but*, *hence*, *in fact*, *therefore*, ...)

# Normalization

**Repetitions** The frequency of content words (words tagged as nouns, verbs, adjectives or adverbs) that occur more than once in a chunk

**Example (Repetitions, Ben-Ari (1998))**

O *Manchmal denke ich: Haben Sie vielleicht ein kaltes Herz? ... Haben Sie wirklich ein kaltes Herz?*

T *Je pense quelquefois: aurait-elle un cœur insensible? ... Avez-vous vraiment un cœur de glace?*

**Contractions** Ratio of contracted forms to their counterpart full form

**Average PMI** The average PMI (Church and Hanks, 1990) of all bigrams in the chunk

**Threshold PMI** Number of bigrams in the chunk whose PMI is above 0

# Normalization

**Repetitions** The frequency of content words (words tagged as nouns, verbs, adjectives or adverbs) that occur more than once in a chunk

## Example (Repetitions, Ben-Ari (1998))

O *Manchmal denke ich: Haben Sie vielleicht ein kaltes Herz? ... Haben Sie wirklich ein kaltes Herz?*

T *Je pense quelquefois: aurait-elle un cœur insensible? ... Avez-vous vraiment un cœur de glace?*

**Contractions** Ratio of contracted forms to their counterpart full form

**Average PMI** The average PMI (Church and Hanks, 1990) of all bigrams in the chunk

**Threshold PMI** Number of bigrams in the chunk whose PMI is above 0

# Normalization

**Repetitions** The frequency of content words (words tagged as nouns, verbs, adjectives or adverbs) that occur more than once in a chunk

## Example (Repetitions, Ben-Ari (1998))

O *Manchmal denke ich: Haben Sie vielleicht ein kaltes Herz? ... Haben Sie wirklich ein kaltes Herz?*

T *Je pense quelquefois: aurait-elle un cœur insensible? ... Avez-vous vraiment un cœur de glace?*

**Contractions** Ratio of contracted forms to their counterpart full form

**Average PMI** The average PMI (Church and Hanks, 1990) of all bigrams in the chunk

**Threshold PMI** Number of bigrams in the chunk whose PMI is above 0

# Normalization

**Repetitions** The frequency of content words (words tagged as nouns, verbs, adjectives or adverbs) that occur more than once in a chunk

## Example (Repetitions, Ben-Ari (1998))

O *Manchmal denke ich: Haben Sie vielleicht ein kaltes Herz? ... Haben Sie wirklich ein kaltes Herz?*

T *Je pense quelquefois: aurait-elle un cœur insensible? ... Avez-vous vraiment un cœur de glace?*

**Contractions** Ratio of contracted forms to their counterpart full form

**Average PMI** The average PMI (Church and Hanks, 1990) of all bigrams in the chunk

**Threshold PMI** Number of bigrams in the chunk whose PMI is above 0

# Normalization

**Repetitions** The frequency of content words (words tagged as nouns, verbs, adjectives or adverbs) that occur more than once in a chunk

**Example (Repetitions, Ben-Ari (1998))**

O *Manchmal denke ich: Haben Sie vielleicht ein kaltes Herz? ... Haben Sie wirklich ein kaltes Herz?*

T *Je pense quelquefois: aurait-elle un cœur insensible? ... Avez-vous vraiment un cœur de glace?*

**Contractions** Ratio of contracted forms to their counterpart full form

**Average PMI** The average PMI (Church and Hanks, 1990) of all bigrams in the chunk

**Threshold PMI** Number of bigrams in the chunk whose PMI is above 0

# Normalization

**Repetitions** The frequency of content words (words tagged as nouns, verbs, adjectives or adverbs) that occur more than once in a chunk

**Example (Repetitions, Ben-Ari (1998))**

O *Manchmal denke ich: Haben Sie vielleicht ein kaltes Herz? ... Haben Sie wirklich ein kaltes Herz?*

T *Je pense quelquefois: aurait-elle un cœur insensible? ... Avez-vous vraiment un cœur de glace?*

**Contractions** Ratio of contracted forms to their counterpart full form

**Average PMI** The average PMI (Church and Hanks, 1990) of all bigrams in the chunk

**Threshold PMI** Number of bigrams in the chunk whose PMI is above 0

# Normalization

**Repetitions** The frequency of content words (words tagged as nouns, verbs, adjectives or adverbs) that occur more than once in a chunk

**Example (Repetitions, Ben-Ari (1998))**

O *Manchmal denke ich: Haben Sie vielleicht ein kaltes Herz? ... Haben Sie wirklich ein kaltes Herz?*

T *Je pense quelquefois: aurait-elle un cœur insensible? ... Avez-vous vraiment un cœur de glace?*

**Contractions** Ratio of contracted forms to their counterpart full form

**Average PMI** The average PMI (Church and Hanks, 1990) of all bigrams in the chunk

**Threshold PMI** Number of bigrams in the chunk whose PMI is above 0

# Interference

- Fingerprints of the source text, "source language shining through" (Teich, 2003)
- Not necessarily a mark of bad translation! Rather, a different distribution of elements in T and O
- Positive vs. negative interference

# Interference

- Fingerprints of the source text, "source language shining through" (Teich, 2003)
- Not necessarily a mark of bad translation! Rather, a different distribution of elements in T and O
- Positive vs. negative interference

## Example (Interference)

# Interference

- Fingerprints of the source text, "source language shining through" (Teich, 2003)
- Not necessarily a mark of bad translation! Rather, a different distribution of elements in T and O
- Positive vs. negative interference

**Example (Interference)**

Positive: *doch* is under-represented in translation to German

Negative: *One is* is over-represented in translations from German to English (triggered by *Man ist*)

# Interference

- Fingerprints of the source text, "source language shining through" (Teich, 2003)
- Not necessarily a mark of bad translation! Rather, a different distribution of elements in T and O
- Positive vs. negative interference

## Example (Interference)

Positive *doch* is under-represented in translation to German

Negative *One is* is over-represented in translations from German to English (triggered by *Man ist*)

# Interference

- Fingerprints of the source text, "source language shining through" (Teich, 2003)
- Not necessarily a mark of bad translation! Rather, a different distribution of elements in T and O
- Positive vs. negative interference

## Example (Interference)

Positive *doch* is under-represented in translation to German

Negative *One is* is over-represented in translations from German to English (triggered by *Man ist*)

# Interference

- Fingerprints of the source text, "source language shining through" (Teich, 2003)
- Not necessarily a mark of bad translation! Rather, a different distribution of elements in T and O
- Positive vs. negative interference

## Example (Interference)

Positive *doch* is under-represented in translation to German

Negative *One is* is over-represented in translations from German to English (triggered by *Man ist*)

# Interference

**POS *n*-grams** Unigrams, bigrams and trigrams of POS tags, modeling variations in syntactic structure

**Character *n*-grams** Unigrams, bigrams and trigrams of characters, modeling shallow morphology

**Prefixes and suffixes** The number of words in a chunk that begin or end with each member of a list of prefixes/suffixes, respectively

**Contextual function words** The frequencies in the chunk of triplets $\langle w_1, w_2, w_3 \rangle$, where at least two of the elements are functions words, and at most one is a POS tag

**Positional token frequency** The frequency of tokens appearing in the first, second, antepenultimate, penultimate and last positions in a sentence

# Interference

**POS *n*-grams** Unigrams, bigrams and trigrams of POS tags, modeling variations in syntactic structure

**Character *n*-grams** Unigrams, bigrams and trigrams of characters, modeling shallow morphology

**Prefixes and suffixes** The number of words in a chunk that begin or end with each member of a list of prefixes/suffixes, respectively

**Contextual function words** The frequencies in the chunk of triplets $\langle w_1, w_2, w_3 \rangle$, where at least two of the elements are functions words, and at most one is a POS tag

**Positional token frequency** The frequency of tokens appearing in the first, second, antepenultimate, penultimate and last positions in a sentence

# Interference

**POS *n*-grams** Unigrams, bigrams and trigrams of POS tags, modeling variations in syntactic structure

**Character *n*-grams** Unigrams, bigrams and trigrams of characters, modeling shallow morphology

**Prefixes and suffixes** The number of words in a chunk that begin or end with each member of a list of prefixes/suffixes, respectively

**Contextual function words** The frequencies in the chunk of triplets $\langle w_1, w_2, w_3 \rangle$, where at least two of the elements are functions words, and at most one is a POS tag

**Positional token frequency** The frequency of tokens appearing in the first, second, antepenultimate, penultimate and last positions in a sentence

# Interference

**POS *n*-grams** Unigrams, bigrams and trigrams of POS tags, modeling variations in syntactic structure

**Character *n*-grams** Unigrams, bigrams and trigrams of characters, modeling shallow morphology

**Prefixes and suffixes** The number of words in a chunk that begin or end with each member of a list of prefixes/suffixes, respectively

**Contextual function words** The frequencies in the chunk of triplets $\langle w_1, w_2, w_3 \rangle$, where at least two of the elements are functions words, and at most one is a POS tag

**Positional token frequency** The frequency of tokens appearing in the first, second, antepenultimate, penultimate and last positions in a sentence

# Interference

**POS *n*-grams** Unigrams, bigrams and trigrams of POS tags, modeling variations in syntactic structure

**Character *n*-grams** Unigrams, bigrams and trigrams of characters, modeling shallow morphology

**Prefixes and suffixes** The number of words in a chunk that begin or end with each member of a list of prefixes/suffixes, respectively

**Contextual function words** The frequencies in the chunk of triplets $\langle w_1, w_2, w_3 \rangle$, where at least two of the elements are functions words, and at most one is a POS tag

**Positional token frequency** The frequency of tokens appearing in the first, second, antepenultimate, penultimate and last positions in a sentence

# Miscellaneous

**Function words** The list of Koppel and Ordan (2011)

**Pronouns** Frequency of the occurrences of pronouns in the chunk

**Punctuation** ? ! : ; - ( ) [ ] ‘ ’ “ ” / , .

1. The normalized frequency of each punctuation mark in the chunk
2. A non-normalized notion of frequency: $\frac{p}{|\text{chunk}|}$, where $p$ is the total number of punctuations in the chunk

**Ratio of passive forms to all verbs**

**Sanity check** Token unigrams and bigrams

# Miscellaneous

**Function words** The list of Koppel and Ordan (2011)

**Pronouns** Frequency of the occurrences of pronouns in the chunk

**Punctuation** ? ! : ; - ( ) [ ] ‘ ’ “ ” / , .

1. The normalized frequency of each punctuation mark in the chunk
2. A non-normalized notion of frequency: $p/tokens$
3. $\frac{\%p}{p}$, where $p$ is the total number of punctuations in the chunk

**Ratio of passive forms to all verbs**

**Sanity check** Token unigrams and bigrams

# Miscellaneous

**Function words** The list of Koppel and Ordan (2011)

**Pronouns** Frequency of the occurrences of pronouns in the chunk

**Punctuation** *? ! : ; - ( ) [ ] ‘ ’ “ ” / , .*

1. The normalized frequency of each punctuation mark in the chunk
2. A non-normalized notion of frequency: $n/tokens$
3. $n/p$, where $p$ is the total number of punctuations in the chunk

**Ratio of passive forms to all verbs**

**Sanity check** Token unigrams and bigrams

# Miscellaneous

**Function words** The list of Koppel and Ordan (2011)

**Pronouns** Frequency of the occurrences of pronouns in the chunk

**Punctuation** *? ! : ; - ( ) [ ] ‘ ’ “ ” / , .*

1. The normalized frequency of each punctuation mark in the chunk
2. A non-normalized notion of frequency: $n/tokens$
3. $n/p$, where $p$ is the total number of punctuations in the chunk

**Ratio of passive forms to all verbs**

**Sanity check** Token unigrams and bigrams

# Miscellaneous

**Function words** The list of Koppel and Ordan (2011)

**Pronouns** Frequency of the occurrences of pronouns in the chunk

**Punctuation** *? ! : ; - ( ) [ ] ‘ ’ “ ” / , .*

1. The normalized frequency of each punctuation mark in the chunk
2. A non-normalized notion of frequency: $n/tokens$
3. $n/p$, where $p$ is the total number of punctuations in the chunk

**Ratio of passive forms to all verbs**

**Sanity check** Token unigrams and bigrams

# Miscellaneous

**Function words** The list of Koppel and Ordan (2011)

**Pronouns** Frequency of the occurrences of pronouns in the chunk

**Punctuation** *? ! : ; - ( ) [ ] ‘ ’ “ ” / , .*

1. The normalized frequency of each punctuation mark in the chunk
2. A non-normalized notion of frequency: $n/tokens$
3. $n/p$, where $p$ is the total number of punctuations in the chunk

**Ratio of passive forms to all verbs**

**Sanity check** Token unigrams and bigrams

# Miscellaneous

**Function words** The list of Koppel and Ordan (2011)

**Pronouns** Frequency of the occurrences of pronouns in the chunk

**Punctuation** *? ! : ; - ( ) [ ] ‘ ’ “ ” / , .*

1. The normalized frequency of each punctuation mark in the chunk
2. A non-normalized notion of frequency: $n/tokens$
3. $n/p$, where $p$ is the total number of punctuations in the chunk

**Ratio of passive forms to all verbs**

**Sanity check** Token unigrams and bigrams

# Miscellaneous

**Function words** The list of Koppel and Ordan (2011)

**Pronouns** Frequency of the occurrences of pronouns in the chunk

**Punctuation** *? ! : ; - ( ) [ ] ‘ ’ “ ” / , .*

1. The normalized frequency of each punctuation mark in the chunk
2. A non-normalized notion of frequency: $n/tokens$
3. $n/p$, where $p$ is the total number of punctuations in the chunk

**Ratio of passive forms to all verbs**

**Sanity check** Token unigrams and bigrams

# Outline

# Results: Sanity Check

| Category | Feature | Accuracy (%) |
|---|---|---|
| Sanity | Token unigrams | 100 |
| | Token bigrams | 100 |

# Results: Simplification

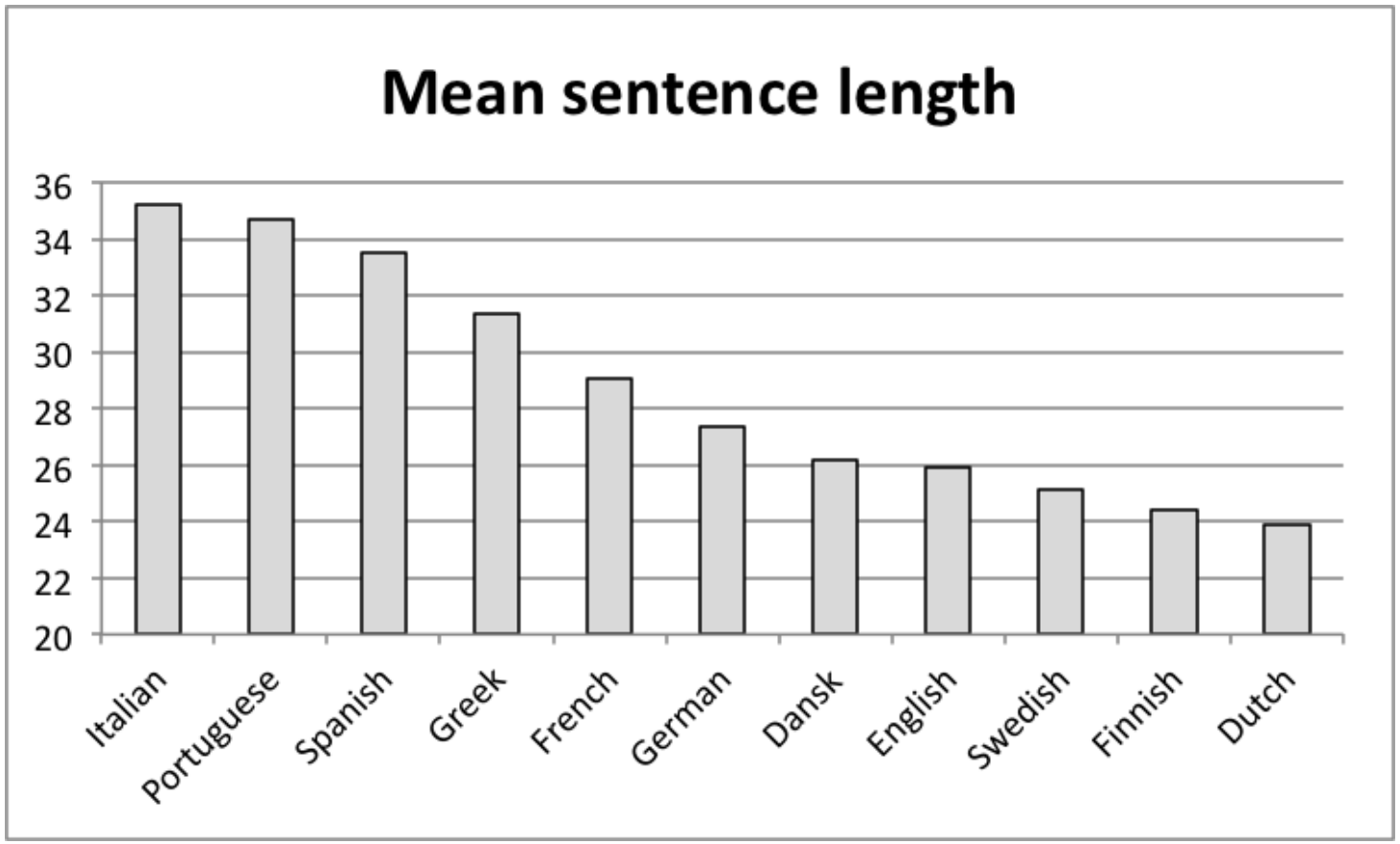
| Category | Feature | Accuracy (%) |
|---|---|---|
| Simplification | TTR (1) | 72 |
| | TTR (2) | 72 |
| | TTR (3) | 76 |
| | Mean word rank (1) | 69 |
| | Mean word rank (2) | 77 |
| | $N$ most frequent words | 64 |
| | Mean word length | 66 |
| | Syllable ratio | 61 |
| | Lexical density | 53 |
| | Mean sentence length | 65 |

# Analysis: Simplification

- Lexical density fails altogether to predict the status of a text, being nearly on chance level (53% accuracy)
- The first two TTR measures perform relatively well (72%), and the indirect measures of lexical variety (mean word length, 66% and syllable ratio, 61%) are above chance level
- Mean word rank (77%) is closely related to the feature studied by Laviosa (1998) (*n* top words) with two differences: our list of frequent items is much larger, and we generate the frequency list not from the corpora under study but rather from an external much larger reference corpus
- In contrast, the design that follows Laviosa (1998) more strictly (*n* most frequent words) has a lower predictive power (64%)
- While mean sentence length is much above chance level (65%), the results are contrary to common assumptions in Translation Studies

# Simplification

## Example (Mean sentence length per 'language')



# Results: Explicitation

| Category | Feature | Accuracy (%) |
|---|---|---|
| Explicitation | Cohesive Markers | 81 |
| | Explicit naming | 58 |
| | Single naming | 56 |
| | Mean multiple naming | 54 |

# Analysis: Explicitation

- Explicit naming, single naming and mean multiple naming do not exceed 58% accuracy
- On the other hand, the classifier that uses 40 cohesive markers (Blum-Kulka, 1986; Koppel and Ordan, 2011) achieves 81% accuracy
- Such cohesive markers are far more frequent in T than in O
- For example, *moreover* is used 17.5 times more frequently in T than in O; *thus* 4 times more frequently; and *besides* 3.8 times

# Results: Normalization

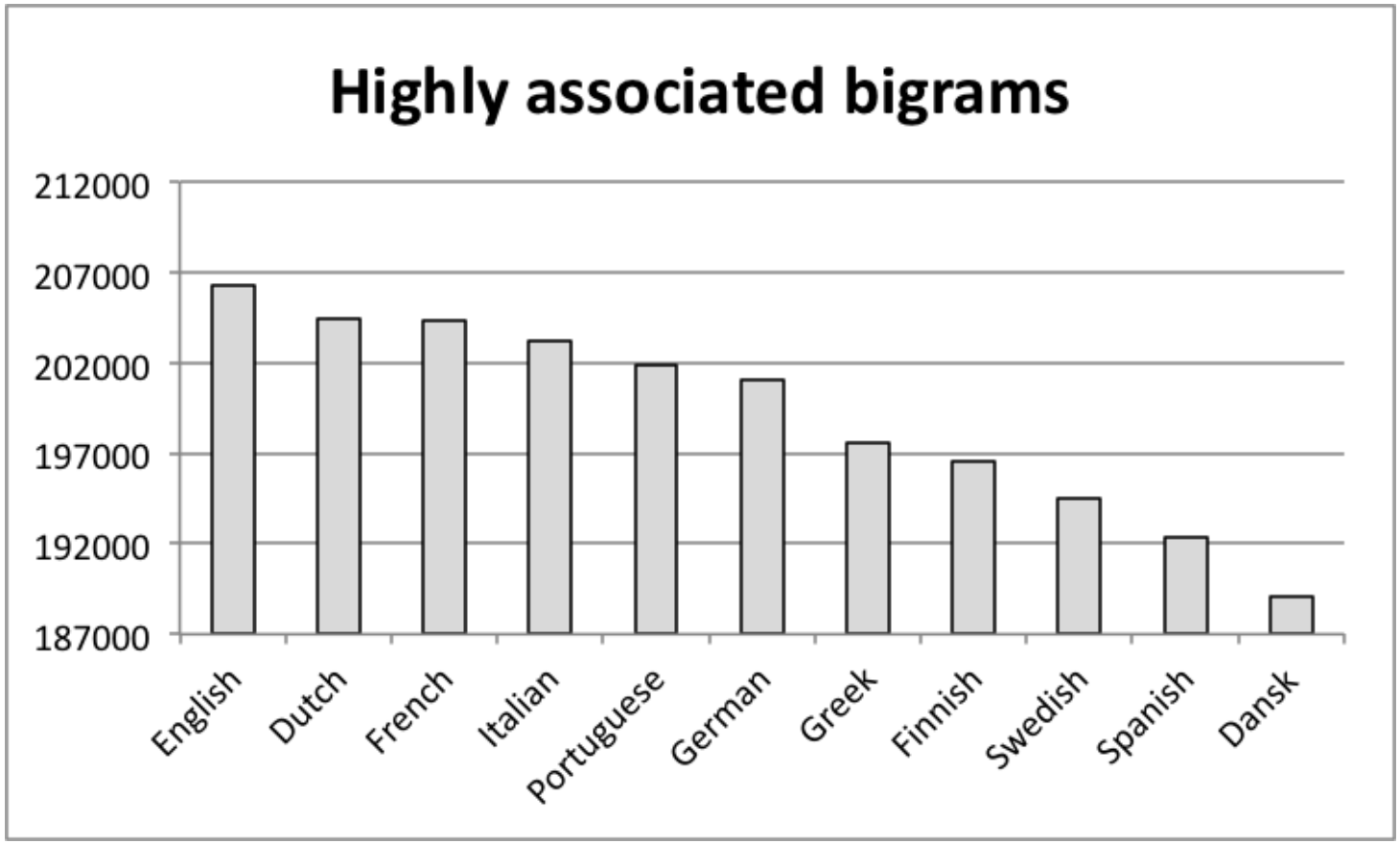
| Category | Feature | Accuracy (%) |
|---|---|---|
| Normalization | Repetitions | 55 |
| | Contractions | 50 |
| | Average PMI | 52 |
| | Threshold PMI | 66 |

# Analysis: Normalization

- None of these features perform very well
- Repetitions and contractions are rare in EUROPARL; the corpus may not be suited for studying these phenomena
- Threshold PMI, designed to pick on highly associated words and therefore attesting to many fixed expressions, performs considerably better, at 66% accuracy, but contrary to the hypothesis (Kenny, 2001)
- English has far more highly associated bigrams than translations: O has *stand idly*, *stand firm*, *stand trial*, etc.; conversely, T includes poorly associated bigrams such as *stand unamended*

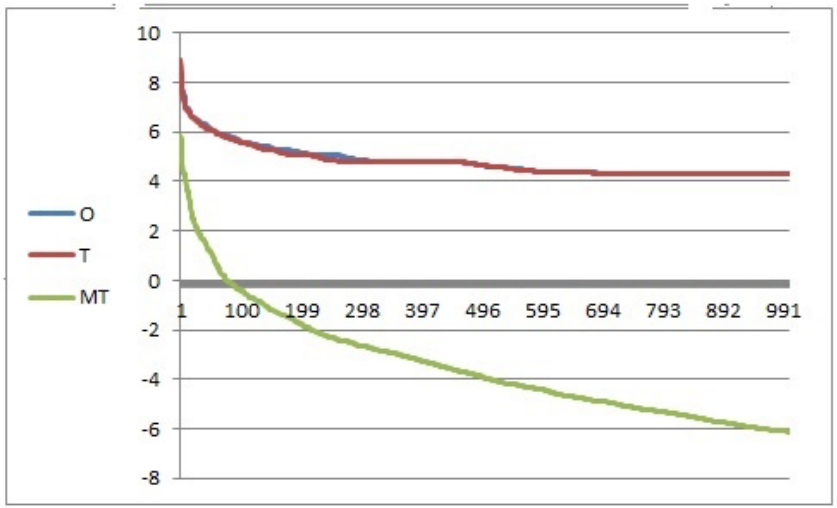
# Normalization

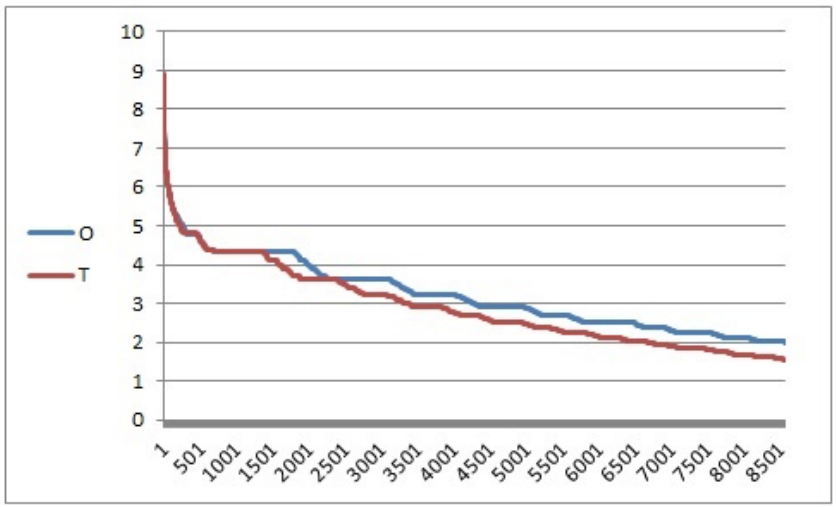
## Example (Number of bigrams whose PMI is above threshold according to 'language')



# Normalization

## Example (Translated language is less predictable)



# Results: Interference

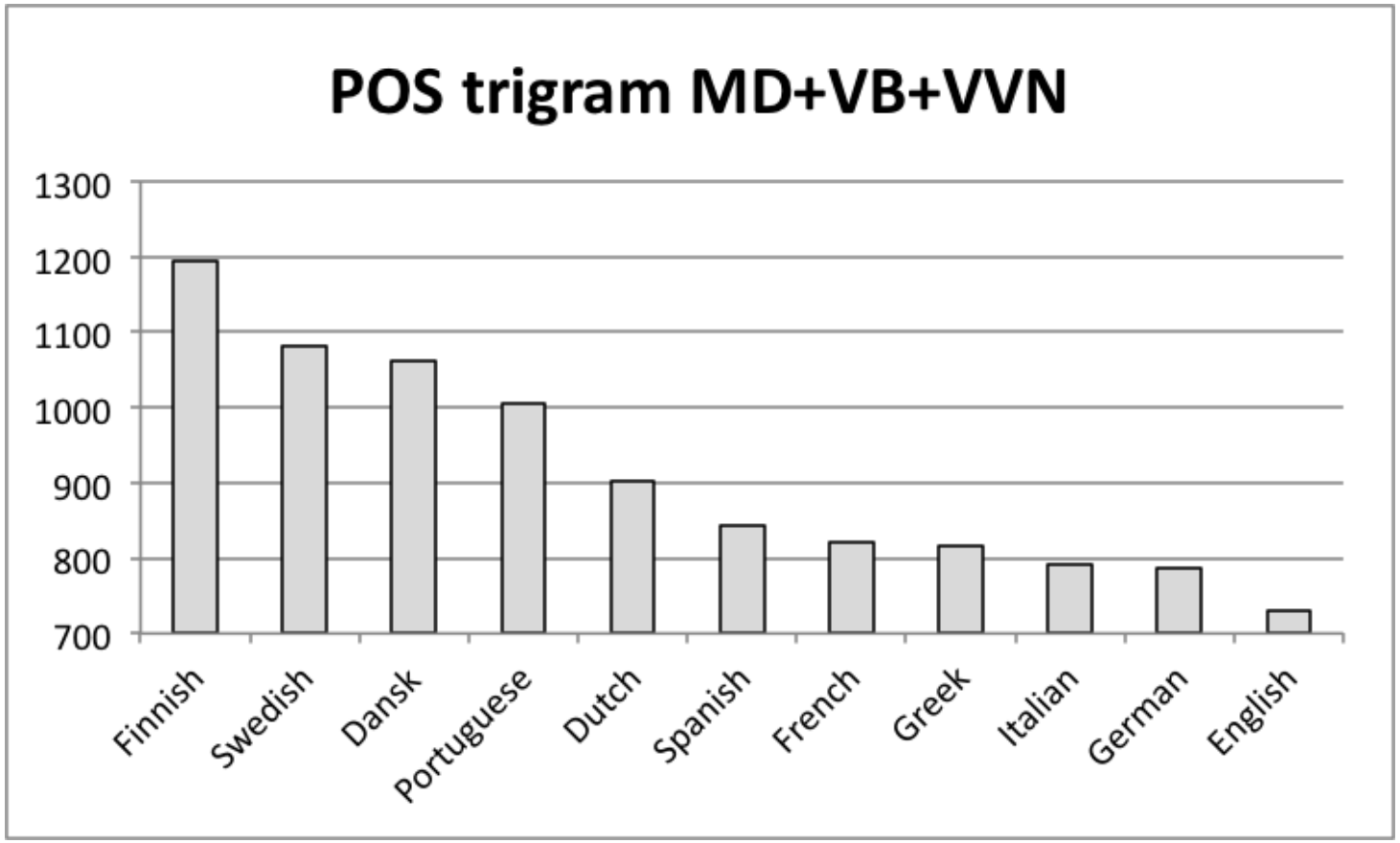
| Category | Feature | Accuracy (%) |
|---|---|---|
| Interference | POS unigrams | 90 |
| | POS bigrams | 97 |
| | POS trigrams | 98 |
| | Character unigrams | 85 |
| | Character bigrams | 98 |
| | Character trigrams | 100 |
| | Prefixes and suffixes | 80 |
| | Contextual function words | 100 |
| | Positional token frequency | 97 |

# Analysis: Interference

- The interference-based classifiers are the best performing ones
- Interference is the most robust phenomenon typifying translations
- The character *n*-gram findings are consistent with Popescu (2011): they catch on both affixes and function words
- For example, typical trigrams in O are *-ion* and *all* whereas typical to T are *-ble* and *the*
- Part-of-speech trigrams is an extremely cheap and efficient classifier
- For example, *MD+VB+VBN*, e.g., *must be taken*, *should be given*, *can be used*

# Interference

## Example (The average number of the POS trigram modal + verb base form + participle in O and ten Ts)



# Analysis: Interference

- Positional token frequency yields 97% accuracy
- The second most prominent feature typifying O is sentences opening with the word 'But'; there are 2.25 times more cases of such sentences in O
- A long prescriptive tradition in English forbids writers to open a sentence with 'But', and translators are conservative in their lexical choices (Kenny, 2001)
- This is in fact **standardization** rather than interference

# Analysis: Interference

- Some interference features are coarse and may reflect corpus-dependent characteristics
- For example, some of the top character *n*-gram features in O include sequences that are 'illegal' in English and obviously stem from foreign names: *H**aa**rder* and *M**aa**t* or *Ga**zp**ron*
- To offset this problem we use only the top 300 features in several of the classifiers, with a minor effect on the results

# Results: Reduced Parameter Space

## (300 most frequent features)

| Category | Feature | Accuracy |
|---|---|---|
| Interference | POS bigrams | 96 |
| | POS trigrams | 96 |
| | Character bigrams | 95 |
| | Character trigrams | 96 |
| | Positional token frequency | 93 |

# Results: Miscellaneous

| Category | Feature | Accuracy (%) |
|---|---|---|
| Miscellaneous | Function words | 96 |
| | Punctuation (1) | 81 |
| | Punctuation (2) | 85 |
| | Punctuation (3) | 80 |
| | Pronouns | 77 |
| | Ratio of passive forms to all verbs | 65 |

# Analysis: Miscellaneous

- The function words classifier replicates Koppel and Ordan (2011); despite the good performance (96% accuracy) it is not very meaningful theoretically
- Subject pronouns, like *I*, *he* and *she* are prominent indicators of O, whereas virtually all reflexive pronouns (such as *itself*, *himself*, *yourself*) typify T
- T has about 1.15 times more passive verbs, but it is highly dependent on the source language
- Punctuation marks are good predictors
- The mark '.' (actually indicating sentence length) is a strong feature of O; ',' is a strong marker of T
- In fact, these two features alone yield 79% accuracy
- Parentheses are very typical to T, indicating explicitation
- Translations from German use many more exclamation marks, 2.76 times more than original English!!!

# Analysis: Punctuation

## Example (Punctuation marks across O and T)

| Mark | Frequency O | Frequency T | Ratio | LL | Weight |
|---|---|---|---|---|---|
| , | 37.83 | 49.79 | 0.76 | T | T1 |
| ( | 0.42 | 0.72 | 0.58 | T | T2 |
| ' | 1.94 | 2.53 | 0.77 | T | T3 |
| ) | 0.40 | 0.72 | 0.56 | T | T4 |
| / | 0.30 | 0.30 | 1.00 | — | — |
| [ | 0.01 | 0.02 | 0.45 | T | — |
| ] | 0.01 | 0.02 | 0.44 | T | — |
| ” | 0.33 | 0.22 | 1.46 | O | O7 |
| ! | 0.22 | 0.17 | 1.25 | O | O6 |
| . | 38.20 | 34.60 | 1.10 | O | O5 |
| : | 1.20 | 1.17 | 1.03 | — | O4 |
| ; | 0.84 | 0.83 | 1.01 | — | O3 |
| ? | 1.57 | 1.11 | 1.41 | O | O2 |
| - | 2.68 | 2.25 | 1.19 | O | O1 |

# Outline

# Supervised Classification

- Inherently dependent on data annotated with the translation direction
- May not be generalized to unseen (related or unrelated) domains: modality (written vs. spoken), register, genre, date, etc.

# Supervised Classification

- Inherently dependent on data annotated with the translation direction
- May not be generalized to unseen (related or unrelated) domains: modality (written vs. spoken), register, genre, date, etc.

# Datasets

- Europarl, the proceedings of the European Parliament, between the years 2001-2006
- the Canadian Hansard, transcripts of the Canadian Parliament, spanning years 2001-2009
- literary classics written (or translated) mainly in the 19th century
- transcripts of TED and TEDx talks

# Datasets

- Europarl, the proceedings of the European Parliament, between the years 2001-2006
- the Canadian Hansard, transcripts of the Canadian Parliament, spanning years 2001-2009
- literary classics written (or translated) mainly in the 19th century
- transcripts of TED and TEDx talks

# Datasets

- Europarl, the proceedings of the European Parliament, between the years 2001-2006
- the Canadian Hansard, transcripts of the Canadian Parliament, spanning years 2001-2009
- literary classics written (or translated) mainly in the 19th century
- transcripts of TED and TEDx talks

# Datasets

- Europarl, the proceedings of the European Parliament, between the years 2001-2006
- the Canadian Hansard, transcripts of the Canadian Parliament, spanning years 2001-2009
- literary classics written (or translated) mainly in the 19th century
- transcripts of TED and TEDx talks

# Supervised Classification

| feature / corpus | EUR | HAN | LIT | TED |
|---|---|---|---|---|
| FW | 96.3 | 98.1 | 97.3 | 97.7 |
| char-trigrams | 98.8 | 97.1 | 99.5 | 100.0 |
| POS-trigrams | 98.5 | 97.2 | 98.7 | 92.0 |
| contextual FW | 95.2 | 96.8 | 94.1 | 86.3 |
| cohesive markers | 83.6 | 86.9 | 78.6 | 81.8 |

# Pairwise Cross-domain Classification

## Using function words

| train / test | EUR | HAN | LIT | X-validation |
|---|---|---|---|---|
| EUR | | 60.8 | 56.2 | 96.3 |
| HAN | 59.7 | | 58.7 | 98.1 |
| LIT | 64.3 | 61.5 | | 97.3 |

# Leave-one-out Cross-domain Classification

## Using function words

| train / test | EUR | HAN | LIT | X-validation |
|---|---|---|---|---|
| EUR + HAN | | | 63.8 | 94.0 |
| EUR + LIT | | 64.1 | | 92.9 |
| HAN + LIT | 59.8 | | | 96.0 |

# Clustering

## Assuming gold labels

| feature / corpus | EUR | HAN | LIT | TED |
|---|---|---|---|---|
| FW | 88.6 | 88.9 | **78.8** | **87.5** |
| char-trigrams | 72.1 | 63.8 | 70.3 | 78.6 |
| POS-trigrams | **96.9** | 76.0 | 70.7 | 76.1 |
| contextual FW | 92.9 | **93.2** | 68.2 | 67.0 |
| cohesive markers | 63.1 | 81.2 | 67.1 | 63.0 |

# Cluster Labeling

- Labeling determines which of the clusters is O and which is T
- Clustering can divide observations into classes but cannot label those classes
- Let $O_m$ (O-markers) denote a set of function words that tend to be associated with O; $T_m$ (T-markers) is a set of words typical of T
- Create unigram language models of O and T:

$$p(w \mid O_m) = \frac{tf(w)+\varepsilon}{|O_m|+\varepsilon \times |V|}$$
$$p(w \mid T_m) = \frac{tf(w)+\varepsilon}{|T_m|+\varepsilon \times |V|}$$

- The similarity between a class $X$ (either O or T) and a cluster $C_i$ is determined using the Jensen-Shannon divergence (JSD) (Lin, 1991)

$$D_{JS}(X, C_i) = \sqrt[2]{JSD(P_X || P_{C_i})}$$

# Cluster Labeling

- Labeling determines which of the clusters is O and which is T
- Clustering can divide observations into classes but cannot label those classes
- Let $O_m$ (O-markers) denote a set of function words that tend to be associated with O; $T_m$ (T-markers) is a set of words typical of T
- Create unigram language models of O and T:

$$p(w \mid O_m) = \frac{tf(w)+\varepsilon}{|O_m|+\varepsilon \times |V|}$$
$$p(w \mid T_m) = \frac{tf(w)+\varepsilon}{|T_m|+\varepsilon \times |V|}$$

- The similarity between a class $X$ (either O or T) and a cluster $C_i$ is determined using the Jensen-Shannon divergence (JSD) (Lin, 1991)

$$D_{JS}(X, C_i) = \sqrt[2]{JSD(P_X || P_{C_i})}$$

# Cluster Labeling

- Labeling determines which of the clusters is O and which is T
- Clustering can divide observations into classes but cannot label those classes
- Let $O_m$ (O-markers) denote a set of function words that tend to be associated with O; $T_m$ (T-markers) is a set of words typical of T
- Create unigram language models of O and T:

$$p(w \mid O_m) = \frac{tf(w)+\varepsilon}{|O_m|+\varepsilon \times |V|}$$
$$p(w \mid T_m) = \frac{tf(w)+\varepsilon}{|T_m|+\varepsilon \times |V|}$$

- The similarity between a class $X$ (either O or T) and a cluster $C_i$ is determined using the Jensen-Shannon divergence (JSD) (Lin, 1991)

$$D_{JS}(X, C_i) = \sqrt[2]{JSD(P_X || P_{C_i})}$$

# Cluster Labeling

- Labeling determines which of the clusters is O and which is T
- Clustering can divide observations into classes but cannot label those classes
- Let $O_m$ (O-markers) denote a set of function words that tend to be associated with O; $T_m$ (T-markers) is a set of words typical of T
- Create unigram language models of O and T:

$$p(w \mid O_m) = \frac{tf(w)+\varepsilon}{|O_m|+\varepsilon \times |V|}$$
$$p(w \mid T_m) = \frac{tf(w)+\varepsilon}{|T_m|+\varepsilon \times |V|}$$

- The similarity between a class $X$ (either O or T) and a cluster $C_i$ is determined using the Jensen-Shannon divergence (JSD) (Lin, 1991)

$$D_{JS}(X, C_i) = \sqrt[2]{JSD(P_X || P_{C_i})}$$

# Cluster Labeling

- Labeling determines which of the clusters is O and which is T
- Clustering can divide observations into classes but cannot label those classes
- Let $O_m$ (O-markers) denote a set of function words that tend to be associated with O; $T_m$ (T-markers) is a set of words typical of T
- Create unigram language models of O and T:

$$\begin{aligned} p(w \mid O_m) &= \frac{tf(w)+\varepsilon}{|O_m|+\varepsilon \times |V|} \\ p(w \mid T_m) &= \frac{tf(w)+\varepsilon}{|T_m|+\varepsilon \times |V|} \end{aligned}$$

- The similarity between a class $X$ (either O or T) and a cluster $C_i$ is determined using the Jensen-Shannon divergence (JSD) (Lin, 1991)

$$D_{JS}(X, C_i) = \sqrt[2]{JSD(P_X || P_{C_i})}$$

# Cluster Labeling

- The assignment of the label $X$ to the cluster $C_1$ is then supported by both $C_1$'s proximity to the class $X$ and $C_2$'s proximity to the other class:

$$label(C_1) = \begin{cases} \text{“O”} & \text{if } D_{JS}(O, C_1) \times D_{JS}(T, C_2) < \\ & D_{JS}(O, C_2) \times D_{JS}(T, C_1) \\ \text{“T”} & \text{otherwise} \end{cases}$$

  $C_2$ is then assigned the complementary label

- We select O- and T-markers from a random sample of Europarl and Hansard texts, using 600 chunks from each corpus
- Labeling precision is 100% in all clustering experiments
- This facilitates majority voting of several feature sets

# Cluster Labeling

- The assignment of the label $X$ to the cluster $C_1$ is then supported by both $C_1$'s proximity to the class $X$ and $C_2$'s proximity to the other class:

$$label(C_1) = \begin{cases} \text{“O”} & \text{if } D_{JS}(O, C_1) \times D_{JS}(T, C_2) < \\ & D_{JS}(O, C_2) \times D_{JS}(T, C_1) \\ \text{“T”} & \text{otherwise} \end{cases}$$

  $C_2$ is then assigned the complementary label

- We select O- and T-markers from a random sample of Europarl and Hansard texts, using 600 chunks from each corpus
- Labeling precision is 100% in all clustering experiments
- This facilitates majority voting of several feature sets

# Cluster Labeling

- The assignment of the label $X$ to the cluster $C_1$ is then supported by both $C_1$'s proximity to the class $X$ and $C_2$'s proximity to the other class:

$$label(C_1) = \begin{cases} \text{“O”} & \text{if } D_{JS}(O, C_1) \times D_{JS}(T, C_2) < \\ & D_{JS}(O, C_2) \times D_{JS}(T, C_1) \\ \text{“T”} & \text{otherwise} \end{cases}$$

  $C_2$ is then assigned the complementary label
- We select O- and T-markers from a random sample of Europarl and Hansard texts, using 600 chunks from each corpus
- Labeling precision is 100% in all clustering experiments
- This facilitates majority voting of several feature sets

# Cluster Labeling

- The assignment of the label $X$ to the cluster $C_1$ is then supported by both $C_1$'s proximity to the class $X$ and $C_2$'s proximity to the other class:

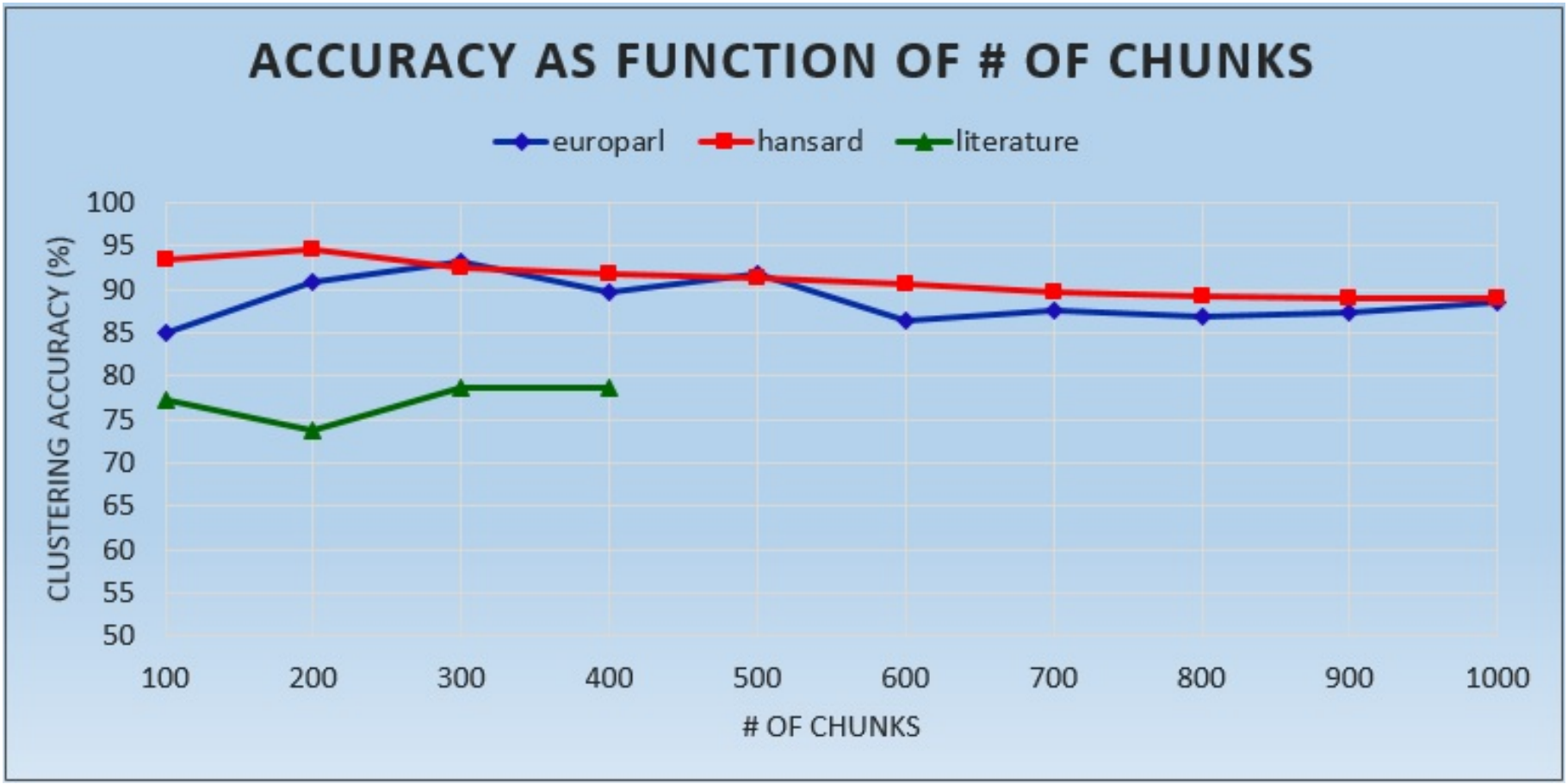
$$label(C_1) = \begin{cases} \text{“O”} & \text{if } D_{JS}(O, C_1) \times D_{JS}(T, C_2) < \\ & D_{JS}(O, C_2) \times D_{JS}(T, C_1) \\ \text{“T”} & \text{otherwise} \end{cases}$$

  $C_2$ is then assigned the complementary label
- We select O- and T-markers from a random sample of Europarl and Hansard texts, using 600 chunks from each corpus
- Labeling precision is 100% in all clustering experiments
- This facilitates majority voting of several feature sets

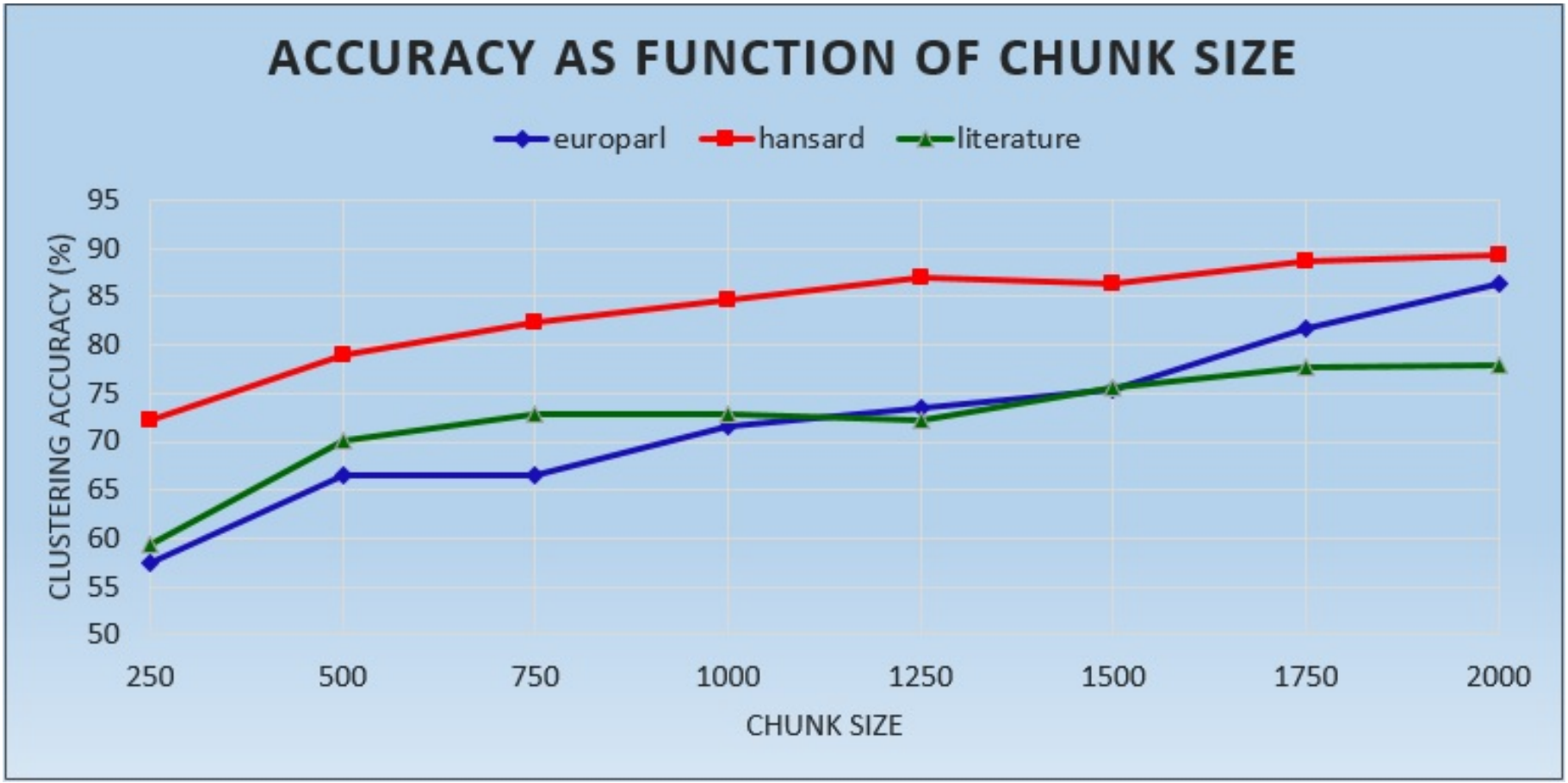
# Clustering Consensus

| method / corpus | EUR | HAN | LIT | TED |
|---|---|---|---|---|
| FW | 88.6 | 88.9 | 78.8 | 87.5 |
| FW<br>char-trigrams<br>POS-trigrams | 91.1 | 86.2 | 78.2 | **90.9** |
| FW<br>POS-trigrams<br>contextual FW | **95.8** | 89.8 | 72.3 | 86.3 |
| FW<br>char-trigrams<br>POS-trigrams<br>contextual FW<br>cohesive markers | 94.1 | **91.0** | **79.2** | 88.6 |

# Sensitivity Analysis



# Sensitivity Analysis



# Mixed-domain Classification

- The in-domain discriminative features of translated texts cannot be easily generalized to other, even related, domains
- Hypothesis: domain-specific features overshadow the features of translationese
- We mix 800 chunks each from Europral and Hansard, yielding 1,600 chunks, half of them O and half T
- Running the clustering algorithm on this dataset yields perfect separation by domain (and chance-level identification of translationese)
- Adding the literature corpus and clustering to three clusters yields the same results

# Mixed-domain Classification

- The in-domain discriminative features of translated texts cannot be easily generalized to other, even related, domains
- Hypothesis: domain-specific features overshadow the features of translationese
- We mix 800 chunks each from Europral and Hansard, yielding 1,600 chunks, half of them O and half T
- Running the clustering algorithm on this dataset yields perfect separation by domain (and chance-level identification of translationese)
- Adding the literature corpus and clustering to three clusters yields the same results

# Mixed-domain Classification

- The in-domain discriminative features of translated texts cannot be easily generalized to other, even related, domains
- Hypothesis: domain-specific features overshadow the features of translationese
- We mix 800 chunks each from Europral and Hansard, yielding 1,600 chunks, half of them O and half T
- Running the clustering algorithm on this dataset yields perfect separation by domain (and chance-level identification of translationese)
- Adding the literature corpus and clustering to three clusters yields the same results

# Mixed-domain Classification

- The in-domain discriminative features of translated texts cannot be easily generalized to other, even related, domains
- Hypothesis: domain-specific features overshadow the features of translationese
- We mix 800 chunks each from Europral and Hansard, yielding 1,600 chunks, half of them O and half T
- Running the clustering algorithm on this dataset yields perfect separation by domain (and chance-level identification of translationese)
- Adding the literature corpus and clustering to three clusters yields the same results

# Mixed-domain Classification

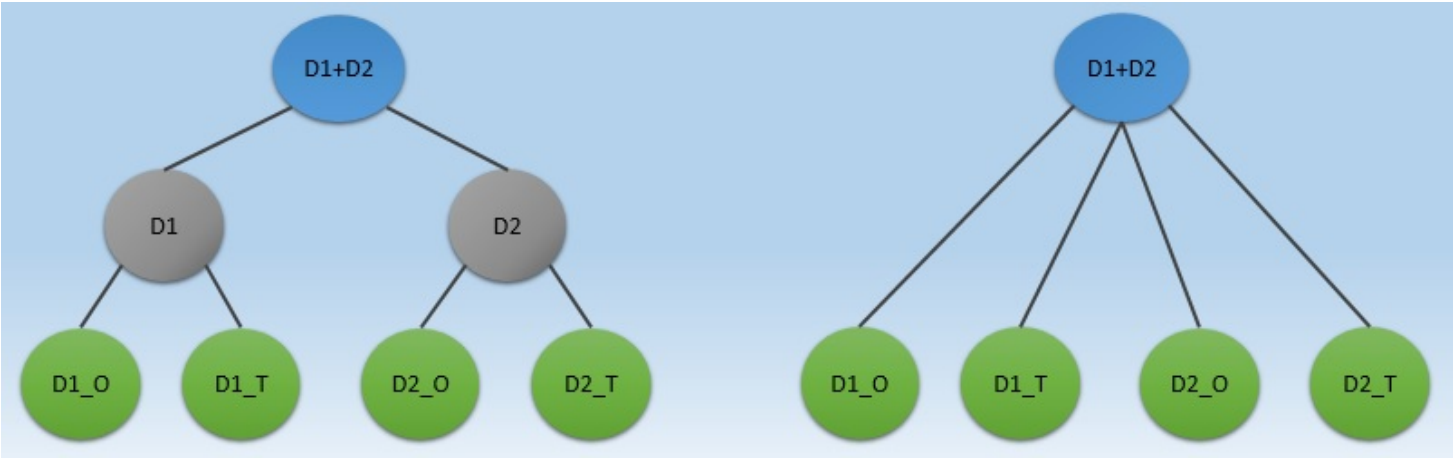
- The in-domain discriminative features of translated texts cannot be easily generalized to other, even related, domains
- Hypothesis: domain-specific features overshadow the features of translationese
- We mix 800 chunks each from Europral and Hansard, yielding 1,600 chunks, half of them O and half T
- Running the clustering algorithm on this dataset yields perfect separation by domain (and chance-level identification of translationese)
- Adding the literature corpus and clustering to three clusters yields the same results

# Mixed-domain Classification

| method / corpus | EUR HAN | EUR LIT | HAN LIT | EUR HAN LIT |
|---|---|---|---|---|
| **KMeans** | | | | |
| accuracy by domain | 93.7 | 99.5 | 99.8 | 92.2 |
| **XMeans** | | | | |
| generated # of clusters | 2 | 2 | 3 | 3 |
| accuracy by domain | 93.6 | 99.5 | – | 92.2 |

# Clustering in a Mixed-domain Setup

- Given a set of text chunks that come from multiple domains, such that some chunks are O and some are T, the task is to classify the texts to O vs. T, **independently of their domain**
- A **two-phase** approach
- A **flat** approach



# Clustering in a Mixed-domain Setup

| method / corpus | EUR HAN | EUR LIT | HAN LIT | EUR HAN LIT |
|---|---|---|---|---|
| Flat | **92.5** | 60.7 | 77.5 | 66.8 |
| Two-phase | 91.3 | **79.4** | **85.3** | **67.5** |

# Outline

# Applications for Machine Translation

- Fundamentals of statistical machine translation (SMT)
- Language models
- Translation models

# Applications for Machine Translation

- Fundamentals of statistical machine translation (SMT)
- Language models
- Translation models

# Applications for Machine Translation

- Fundamentals of statistical machine translation (SMT)
- Language models
- Translation models

# Fundamentals of SMT

- Motivation

> *When I look at an article in Russian, I say, "This is really written in English, but it has been coded in some strange symbols. I shall now proceed to decode."*
>
> *Warren Weaver, 1955*

- The **noisy channel model**
- The best translation $\hat{T}$ of a source sentence $S$ is the target sentence $T$ that maximizes some function combining the **faithfulness** of $(T, S)$ and the **fluency** of $T$

# Fundamentals of SMT

- Motivation

  *When I look at an article in Russian, I say, "This is really written in English, but it has been coded in some strange symbols. I shall now proceed to decode."*

  *Warren Weaver, 1955*

- The **noisy channel model**
- The best translation $\hat{T}$ of a source sentence $S$ is the target sentence $T$ that maximizes some function combining the **faithfulness** of $(T, S)$ and the **fluency** of $T$

# Fundamentals of SMT

- Motivation

  *When I look at an article in Russian, I say, "This is really written in English, but it has been coded in some strange symbols. I shall now proceed to decode."*

  *Warren Weaver, 1955*

- The **noisy channel model**
- The best translation $\hat{T}$ of a source sentence $S$ is the target sentence $T$ that maximizes some function combining the **faithfulness** of $(T, S)$ and the **fluency** of $T$

# Fundamentals of SMT

- Motivation

  *When I look at an article in Russian, I say, "This is really written in English, but it has been coded in some strange symbols. I shall now proceed to decode."*

  *Warren Weaver, 1955*

- The **noisy channel model**
- The best translation $\hat{T}$ of a source sentence $S$ is the target sentence $T$ that maximizes some function combining the **faithfulness** of $(T, S)$ and the **fluency** of $T$

# Fundamentals of SMT

- Standard notation: Translating a **foreign** sentence $F = f_1, \cdots, f_m$ into an **English** sentence $E = e_1, \cdots, e_l$
- The best translation:

$$\begin{aligned} \hat{E} &= \arg\max_E P(E \mid F) \\ &= \arg\max_E \frac{P(F|E) \times P(E)}{P(F)} \\ &= \arg\max_E P(F \mid E) \times P(E) \end{aligned}$$

- The noisy channel thus requires two components: a **translation model** and a **language model**

$$\hat{E} = \arg\max_{E \in English} \underbrace{P(F \mid E)}_{\text{Translation model}} \times \underbrace{P(E)}_{\text{Language model}}$$

# Fundamentals of SMT

- Standard notation: Translating a **foreign** sentence $F = f_1, \cdots, f_m$ into an **English** sentence $E = e_1, \cdots, e_l$
- The best translation:

$$\begin{aligned} \hat{E} &= \arg\max_E P(E \mid F) \\ &= \arg\max_E \frac{P(F|E) \times P(E)}{P(F)} \\ &= \arg\max_E P(F \mid E) \times P(E) \end{aligned}$$

- The noisy channel thus requires two components: a **translation model** and a **language model**

$$\hat{E} = \arg\max_{E \in English} \underbrace{P(F \mid E)}_{Translation\ model} \times \underbrace{P(E)}_{Language\ model}$$

# Fundamentals of SMT

- Standard notation: Translating a **foreign** sentence $F = f_1, \cdots, f_m$ into an **English** sentence $E = e_1, \cdots, e_l$
- The best translation:

$$\begin{aligned} \hat{E} &= \arg\max_E P(E \mid F) \\ &= \arg\max_E \frac{P(F|E) \times P(E)}{P(F)} \\ &= \arg\max_E P(F \mid E) \times P(E) \end{aligned}$$

- The noisy channel thus requires two components: a **translation model** and a **language model**

$$\hat{E} = \arg\max_{E \in English} \underbrace{P(F \mid E)}_{Translation\ model} \times \underbrace{P(E)}_{Language\ model}$$

# Fundamentals of SMT

- A **language model** to estimate $P(E)$ (estimated from a monolingual $E$ corpus)
- A **translation model** to estimate $P(F \mid E)$ (estimated from a bilingual parallel corpus)
- A **decoder** that given $F$ can produce the most probable $E$
- Evaluation: BLEU scores (Papineni et al., 2002)

# Fundamentals of SMT

- A **language model** to estimate $P(E)$ (estimated from a monolingual $E$ corpus)
- A **translation model** to estimate $P(F \mid E)$ (estimated from a bilingual parallel corpus)
- A **decoder** that given $F$ can produce the most probable $E$
- Evaluation: BLEU scores (Papineni et al., 2002)

# Fundamentals of SMT

- A **language model** to estimate $P(E)$ (estimated from a monolingual $E$ corpus)
- A **translation model** to estimate $P(F \mid E)$ (estimated from a bilingual parallel corpus)
- A **decoder** that given $F$ can produce the most probable $E$
- Evaluation: BLEU scores (Papineni et al., 2002)

# Fundamentals of SMT

- A **language model** to estimate $P(E)$ (estimated from a monolingual $E$ corpus)
- A **translation model** to estimate $P(F \mid E)$ (estimated from a bilingual parallel corpus)
- A **decoder** that given $F$ can produce the most probable $E$
- Evaluation: BLEU scores (Papineni et al., 2002)

# Language Models

## Research questions

- Test the fitness of language models compiled from translated texts vs. the fitness of LMs compiled from original texts
- Test the fitness of language models compiled from texts translated from other languages
- Test if language models compiled from translated texts are better for MT than LMs compiled from original texts

# Language Models

## Research questions

- Test the fitness of language models compiled from translated texts vs. the fitness of LMs compiled from original texts
- Test the fitness of language models compiled from texts translated from other languages
- Test if language models compiled from translated texts are better for MT than LMs compiled from original texts

# Language Models

## Research questions

- Test the fitness of language models compiled from translated texts vs. the fitness of LMs compiled from original texts
- Test the fitness of language models compiled from texts translated from other languages
- Test if language models compiled from translated texts are better for MT than LMs compiled from original texts

# Language Models

## Methodology

- The fitness of a language model to a reference corpus is evaluated using **perplexity**

$$PP(LM, w_1\, w_2 \ldots w_N) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P_{LM}(w_i \mid w_1 \ldots w_{i-1})}}$$

- Train SMT systems (Koehn et al., 2007) using different LMs and evaluate their quality on a reference set
- Quality is measured in terms of BLEU scores (Papineni et al., 2002)

# Language Models

## Methodology

- The fitness of a language model to a reference corpus is evaluated using **perplexity**

$$PP(LM, w_1\, w_2 \ldots w_N) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P_{LM}(w_i \mid w_1 \ldots w_{i-1})}}$$

- Train SMT systems (Koehn et al., 2007) using different LMs and evaluate their quality on a reference set
- Quality is measured in terms of BLEU scores (Papineni et al., 2002)

# Language Models

## Methodology

- The fitness of a language model to a reference corpus is evaluated using **perplexity**

$$PP(LM, w_1\, w_2 \ldots w_N) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P_{LM}(w_i \mid w_1 \ldots w_{i-1})}}$$

- Train SMT systems (Koehn et al., 2007) using different LMs and evaluate their quality on a reference set
- Quality is measured in terms of BLEU scores (Papineni et al., 2002)

# Perplexity Results

| German to English translations | | | | |
|---|---|---|---|---|
| **Orig. Lang.** | **1-gram** | **2-gram** | **3-gram** | **4-gram** |
| Mix | 451.50 | 93.00 | 69.36 | 66.47 |
| O-EN | *468.09* | *103.74* | *79.57* | *76.79* |
| T-DE | **443.14** | **88.48** | **64.99** | **62.07** |
| T-FR | 460.98 | 99.90 | 76.23 | 73.38 |
| T-IT | 465.89 | 102.31 | 78.50 | 75.67 |
| T-NL | 457.02 | 97.34 | 73.54 | 70.56 |

# Perplexity Results

| French to English translations | | | | |
|---|---|---|---|---|
| **Orig. Lang.** | **1-gram** | **2-gram** | **3-gram** | **4-gram** |
| Mix | 472.05 | 99.04 | 75.60 | 72.68 |
| O-EN | *500.56* | *115.48* | *91.14* | *88.31* |
| T-DE | 486.78 | 108.50 | 84.39 | 81.41 |
| T-FR | **463.58** | **94.59** | **71.24** | **68.37** |
| T-IT | 476.05 | 102.69 | 79.23 | 76.36 |
| T-NL | 490.09 | 110.67 | 86.61 | 83.55 |

# Perplexity Results

| Italian to English translations | | | | |
|---|---|---|---|---|
| **Orig. Lang.** | **1-gram** | **2-gram** | **3-gram** | **4-gram** |
| Mix | 395.99 | 88.46 | 67.35 | 64.40 |
| O-EN | *415.47* | *99.92* | *79.27* | *76.34* |
| T-DE | 404.64 | 95.22 | 73.73 | 70.85 |
| T-FR | 395.99 | 89.44 | 68.38 | 65.54 |
| T-IT | **384.55** | **81.90** | **60.85** | **57.91** |
| T-NL | 411.58 | 98.78 | 76.98 | 73.94 |

# Perplexity Results

| Dutch to English translations | | | | |
|---|---|---|---|---|
| **Orig. Lang.** | **1-gram** | **2-gram** | **3-gram** | **4-gram** |
| Mix | 434.89 | 90.73 | 69.05 | 66.08 |
| O-EN | *448.11* | 100.17 | *78.23* | *75.46* |
| T-DE | 437.68 | 93.67 | 71.54 | 68.57 |
| T-FR | 445.00 | 97.32 | 75.59 | 72.55 |
| T-IT | 448.11 | *100.19* | 78.06 | 75.19 |
| T-NL | **423.13** | **83.99** | **62.17** | **59.09** |

# Language Models

## Abstraction

- No punctuation
- No named entities
- No nouns
- No words: only parts of speech

# Language Models

## Abstraction

- No punctuation
- No named entities
- No nouns
- No words: only parts of speech

# Language Models

## Abstraction

- No punctuation
- No named entities
- No nouns
- No words: only parts of speech

# Language Models

## Abstraction

- No punctuation
- No named entities
- No nouns
- No words: only parts of speech

# Abstraction Results

| No Punctuation | | |
|---|---|---|
| **Orig. Lang.** | **Perplexity** | **Improvement (%)** |
| MIX | 105.91 | 19.73 |
| O-EN | *131.94* | |
| T-DE | 122.50 | 7.16 |
| T-FR | **99.52** | **24.58** |
| T-IT | 112.71 | 14.58 |
| T-NL | 126.44 | 4.17 |

# Abstraction Results

| Named Entity Abstraction | | |
|---|---|---|
| **Orig. Lang.** | **Perplexity** | **Improvement (%)** |
| MIX | 93.88 | 18.51 |
| O-EN | *115.20* | |
| T-DE | 107.48 | 6.70 |
| T-FR | **88.96** | **22.77** |
| T-IT | 99.17 | 13.91 |
| T-NL | 110.72 | 3.89 |

# Abstraction Results

| Noun Abstraction | | |
|---|---|---|
| **Orig. Lang.** | **Perplexity** | **Improvement (%)** |
| MIX | 36.02 | 11.34 |
| O-EN | *40.62* | |
| T-DE | 38.67 | 4.81 |
| T-FR | **34.75** | **14.46** |
| T-IT | 36.85 | 9.30 |
| T-NL | 39.44 | 2.91 |

# Abstraction Results

| Part-of-speech Abstraction | | |
|---|---|---|
| **Orig. Lang.** | **Perplexity** | **Improvement (%)** |
| MIX | 7.99 | 2.66 |
| O-EN | *8.20* | |
| T-DE | 8.08 | 1.47 |
| T-FR | **7.89** | **3.77** |
| T-IT | 8.00 | 2.47 |
| T-NL | 8.11 | 1.11 |

# Machine Translation Results

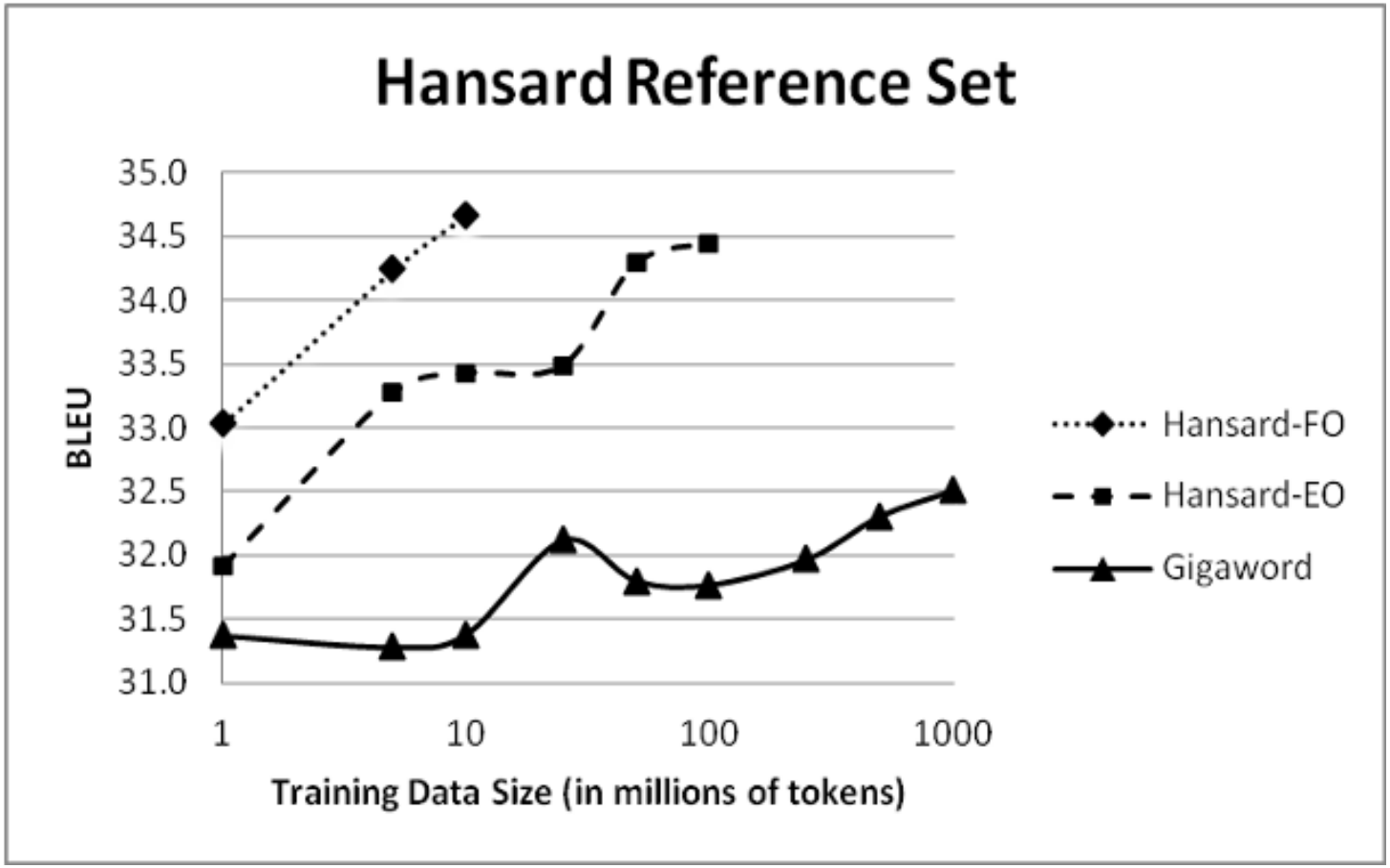
| DE to EN | |
|---|---|
| **LM** | **BLEU** |
| MIX | 21.43 |
| O-EN | *21.10* |
| T-DE | **21.90** |
| T-FR | 21.16 |
| T-IT | 21.29 |
| T-NL | 21.20 |

| FR to EN | |
|---|---|
| **LM** | **BLEU** |
| MIX | 28.67 |
| O-EN | *27.98* |
| T-DE | 28.01 |
| T-FR | **29.14** |
| T-IT | 28.75 |
| T-NL | 28.11 |

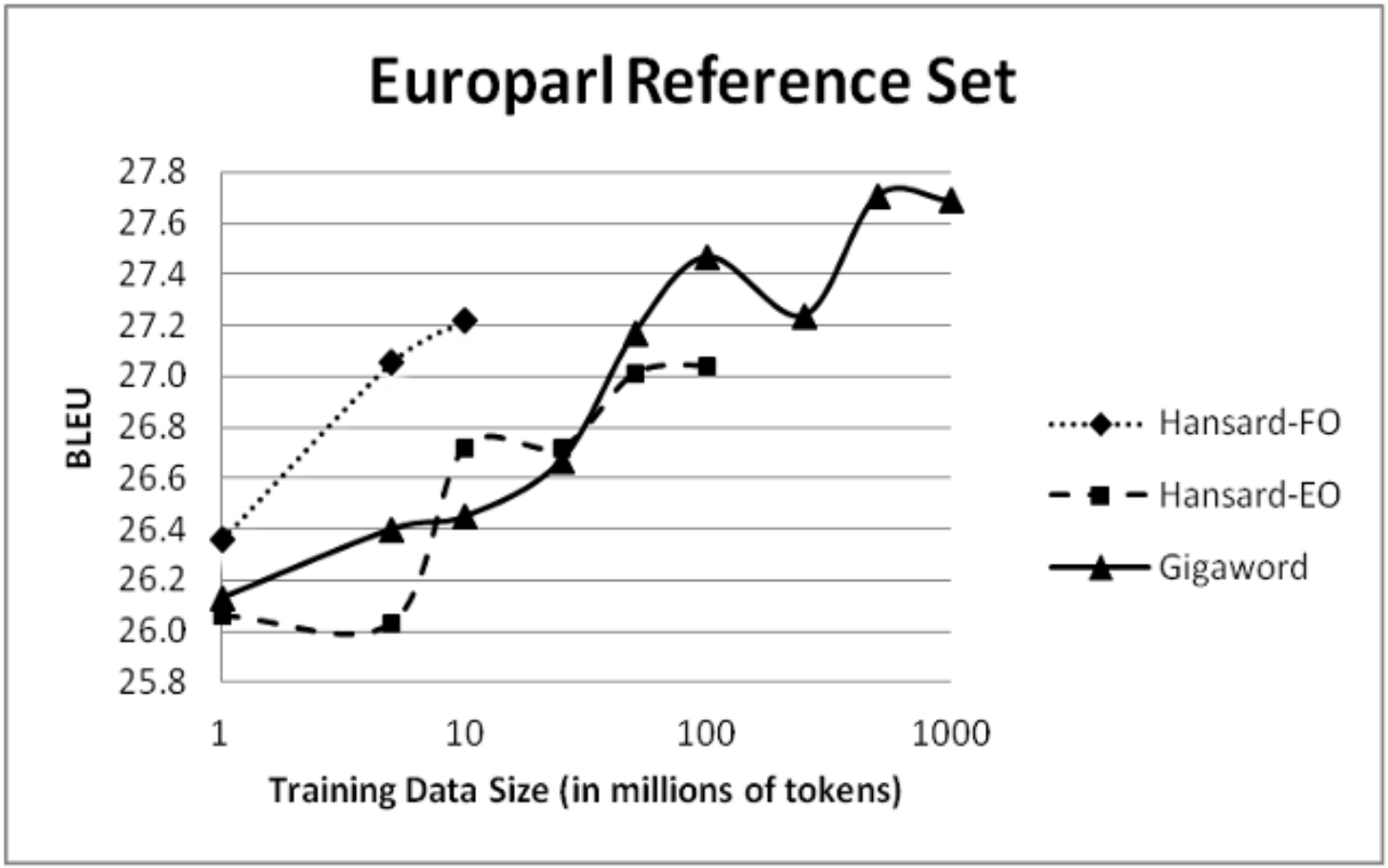
| IT to EN | |
|---|---|
| **LM** | **BLEU** |
| MIX | 25.41 |
| O-EN | 24.69 |
| T-DE | *24.62* |
| T-FR | 25.37 |
| T-IT | **25.96** |
| T-NL | 24.77 |

| NL to EN | |
|---|---|
| **LM** | **BLEU** |
| MIX | 24.20 |
| O-EN | *23.40* |
| T-DE | 24.26 |
| T-FR | 23.56 |
| T-IT | 23.87 |
| T-NL | **24.52** |

# Does Size Matter?



# Does Size Matter?



# Translation Examples

## Cohesive markers

SOURCE *Enfin, ce qui est grave dans le rapport de M. Olivier, c'est qu'il propose une constitution tripotage.*

O *Finally, which is serious in the report of Mr Olivier, is that it proposes a constitution tripotage.*

T *Finally, and this is serious in the report by Mr olivier, it is* **because** *it proposes a constitution tripotage.*

SOURCE *C'est quand même quelque chose de précieux qui a été souligné par tous les membres du conseil européen.*

O *Even when it is something of valuable which has been pointed out by all the members of the European Council.*

T *It is* **nevertheless** *something of a valuable which has been pointed out by all the members of the European Council.*

# Translation Examples

## Verbs

Source *Une telle Europe serait un gage de paix et marquerait le refus de tout nationalisme ethnique.*

O *Such a Europe would be a show of peace and would the rejection of any ethnic nationalism.*

T *Such a Europe would be a show of peace and would* **mark** *the refusal of all ethnic nationalism.*

Source *Votre rapport, madame Sudre, met l'accent, à juste titre, sur la nécessité d'agir dans la durée.*

O *Your report, Mrs Sudre, its emphasis, quite rightly, on the need to act in the long term.*

T *Your report, Mrs Sudre,* **places** *the emphasis, quite rightly, on the need to act in the long term.*

# Translation Examples

## Interference

Source *On ne dit rien non plus sur la responsabilité des fabricants, notamment en grande-bretagne, qui ont été les premiers responsables.*

O *We* **do not say nothing more** *on the responsibility of the manufacturers, particularly in Britain, which were the first responsible.*

T *We* **do not say anything either** *on the responsibility of the manufacturers, particularly in great Britain, who were the first responsible.*

# Translation Models

## Research questions

- Are parallel corpora (manually) translated in the same direction of the MT task better than ones directed in the other direction?
- If corpora consisting of texts (manually) translated in both directions are available, how to build a translation model adapted to the unique properties of the translated text?

# Translation Models

## Research questions

- Are parallel corpora (manually) translated in the same direction of the MT task better than ones directed in the other direction?
- If corpora consisting of texts (manually) translated in both directions are available, how to build a translation model adapted to the unique properties of the translated text?

# Translation Models

## Direction matters!

| Task | $S \rightarrow T$ | $T \rightarrow S$ |
|---|---|---|
| FR-EN | 33.64 | 30.88 |
| EN-FR | 32.11 | 30.35 |
| DE-EN | 26.53 | 23.67 |
| EN-DE | 16.96 | 16.17 |
| IT-EN | 28.70 | 26.84 |
| EN-IT | 23.81 | 21.28 |

# Translation Models

## Direction matters!

| Task: French-to-English | | |
|---|---|---|
| Corpus subset | $S \rightarrow T$ | $T \rightarrow S$ |
| 250K | 34.35 | 31.33 |
| 500K | 35.21 | 32.38 |
| 750K | 36.12 | 32.90 |
| 1M | 35.73 | 33.07 |
| 1.25M | 36.24 | 33.23 |
| 1.5M | 36.43 | 33.73 |

# Translation Models

## Direction matters!

| Task: English-to-French | | |
|---|---|---|
| Corpus subset | $S \rightarrow T$ | $T \rightarrow S$ |
| 250K | 27.74 | 26.58 |
| 500K | 29.15 | 27.19 |
| 750K | 29.43 | 27.63 |
| 1M | 29.94 | 27.88 |
| 1.25M | 30.63 | 27.84 |
| 1.5M | 29.89 | 27.83 |

# Translation Models

## Domain adaptation

**Goal:** Given any bi-text consisting of both $S \rightarrow T$ and $T \rightarrow S$ subsets, improve translation quality by taking advantage of information pertaining to the direction of translation

**Techniques:**

- Union: simple concatenation of corpora
- Two phrase-tables: train a phrase table for each subset and pass both to MOSES
- Phrase table interpolation: using perplexity minimization (Sennrich, 2012)
- Add a feature in the phrase table indicating the direction of translation

# Translation Models

## Domain adaptation

Goal: Given any bi-text consisting of both $S \rightarrow T$ and $T \rightarrow S$ subsets, improve translation quality by taking advantage of information pertaining to the direction of translation

Techniques:

- Union: simple concatenation of corpora
- Two phrase-tables: train a phrase table for each subset and pass both to MOSES
- Phrase table interpolation: using perplexity minimization (Sennrich, 2012)
- Add a feature in the phrase table indicating the direction of translation

# Translation Models

## Domain adaptation

Goal: Given any bi-text consisting of both $S \rightarrow T$ and $T \rightarrow S$ subsets, improve translation quality by taking advantage of information pertaining to the direction of translation

Techniques:

- Union: simple concatenation of corpora
- Two phrase-tables: train a phrase table for each subset and pass both to MOSES
- Phrase table interpolation: using perplexity minimization (Sennrich, 2012)
- Add a feature in the phrase table indicating the direction of translation

# Translation Models

## Domain adaptation

Goal: Given any bi-text consisting of both $S \rightarrow T$ and $T \rightarrow S$ subsets, improve translation quality by taking advantage of information pertaining to the direction of translation

Techniques:

- Union: simple concatenation of corpora
- Two phrase-tables: train a phrase table for each subset and pass both to MOSES
- Phrase table interpolation: using perplexity minimization (Sennrich, 2012)
- Add a feature in the phrase table indicating the direction of translation

# Translation Models

## Domain adaptation

Goal: Given any bi-text consisting of both $S \rightarrow T$ and $T \rightarrow S$ subsets, improve translation quality by taking advantage of information pertaining to the direction of translation

Techniques:

- Union: simple concatenation of corpora
- Two phrase-tables: train a phrase table for each subset and pass both to MOSES
- Phrase table interpolation: using perplexity minimization (Sennrich, 2012)
- Add a feature in the phrase table indicating the direction of translation

# Translation Models

## Domain adaptation

Goal: Given any bi-text consisting of both $S \rightarrow T$ and $T \rightarrow S$ subsets, improve translation quality by taking advantage of information pertaining to the direction of translation

Techniques:

- Union: simple concatenation of corpora
- Two phrase-tables: train a phrase table for each subset and pass both to MOSES
- Phrase table interpolation: using perplexity minimization (Sennrich, 2012)
- Add a feature in the phrase table indicating the direction of translation

# Translation Models

## Adaptation results

| Task: French-to-English | | | |
|---|---|---|---|
| **System** | **MIX** | **MIX-EO** | **MIX-FO** |
| $S \rightarrow T$ | 35.21 | 35.21 | 35.73 |
| UNION | 35.27 | 35.36 | 35.94 |
| PPLMIN-1 | 35.46 | 35.59 | 36.26 |
| PPLMIN-2 | **35.75** | 35.65 | 36.20 |
| CrEnt | 35.54 | 35.45 | **36.75** |
| PplRatio | 35.59 | **35.78** | 36.22 |

# Translation Models

## Adaptation results

| Task: English-to-French | | | |
|---|---|---|---|
| **System** | **MIX** | **MIX-FO** | **MIX-EO** |
| $S \rightarrow T$ | 29.15 | 29.15 | 29.94 |
| UNION | 29.27 | 29.44 | 30.01 |
| PPLMIN-1 | 29.64 | 29.94 | 29.65 |
| PPLMIN-2 | 29.50 | **30.45** | 30.12 |
| CrEnt | 29.47 | 29.45 | **30.44** |
| PplRatio | **29.65** | 29.62 | 30.34 |

# Outline

# The Power of Interference

- Translations to English from $L_1$ differ from translations to English from $L_2$
- Translations from two languages are more similar to each other when the two source languages are closer
- The native language of ESL speakers can be accurately identified by looking at their English texts

# The Power of Interference

- Translations to English from $L_1$ differ from translations to English from $L_2$
- Translations from two languages are more similar to each other when the two source languages are closer
- The native language of ESL speakers can be accurately identified by looking at their English texts

# The Power of Interference

- Translations to English from $L_1$ differ from translations to English from $L_2$
- Translations from two languages are more similar to each other when the two source languages are closer
- The native language of ESL speakers can be accurately identified by looking at their English texts

# Cross-classification

- A classifier is trained to distinguish between O and T translated from $L_1$; it is then used to distinguish O from T translated from $L_2$ (Koppel and Ordan, 2011)

## Example (Train on L1, test on L2)

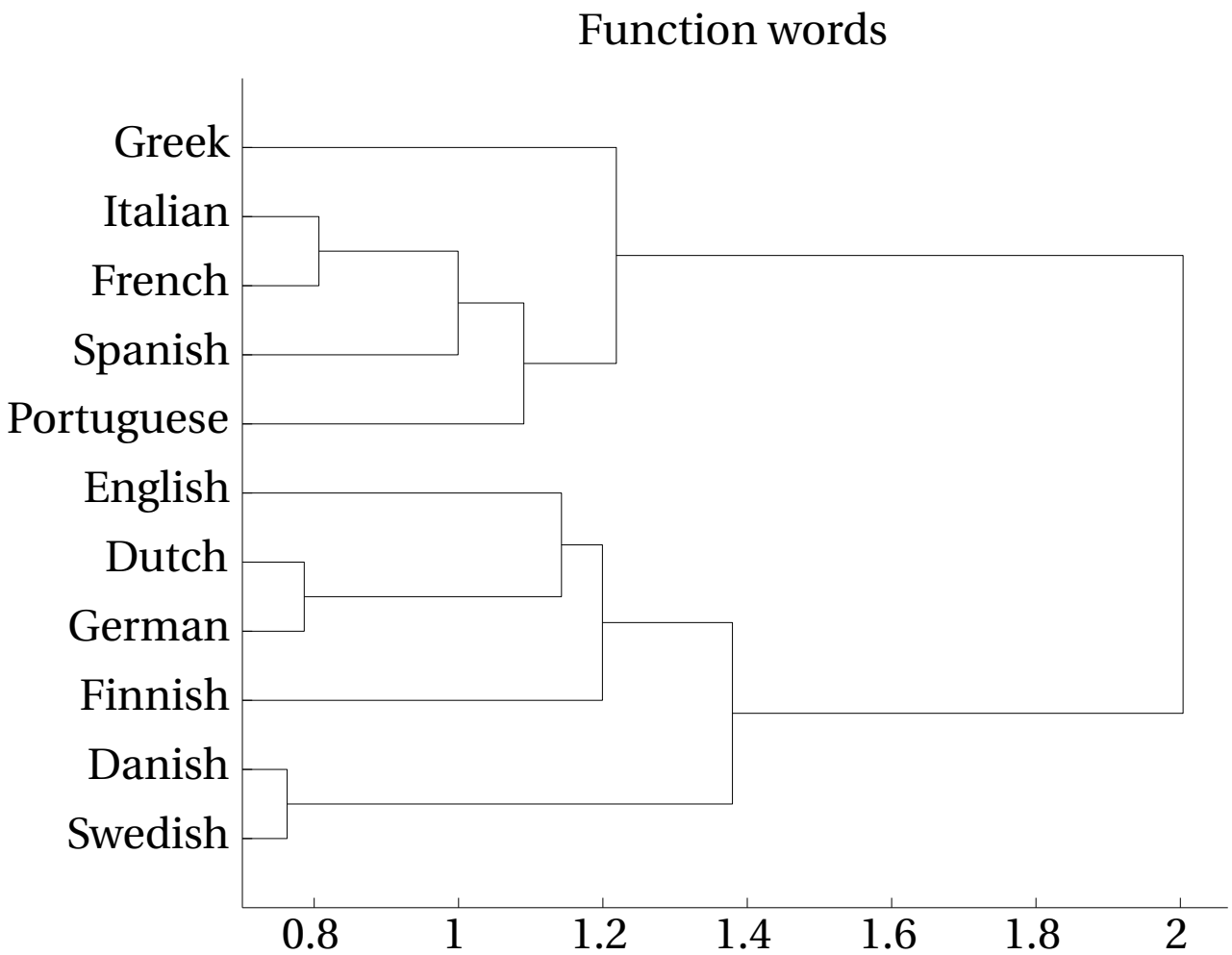
| | IT | FR | ES | DE | FI |
|---|---|---|---|---|---|
| IT | 99 | 93 | 91 | 75 | 63 |
| FR | 90 | 98 | 90 | 83 | 70 |
| ES | 85 | 90 | 98 | 82 | 69 |
| DE | 73 | 74 | 74 | 98 | 70 |
| FI | 58 | 70 | 64 | 81 | 99 |

# Clustering of Translations from Related Languages

- An unsupervised hierarchical clustering algorithm groups together English texts translated from ten European languages
- Texts are represented using feature vectors, similarly to the supervised experiments

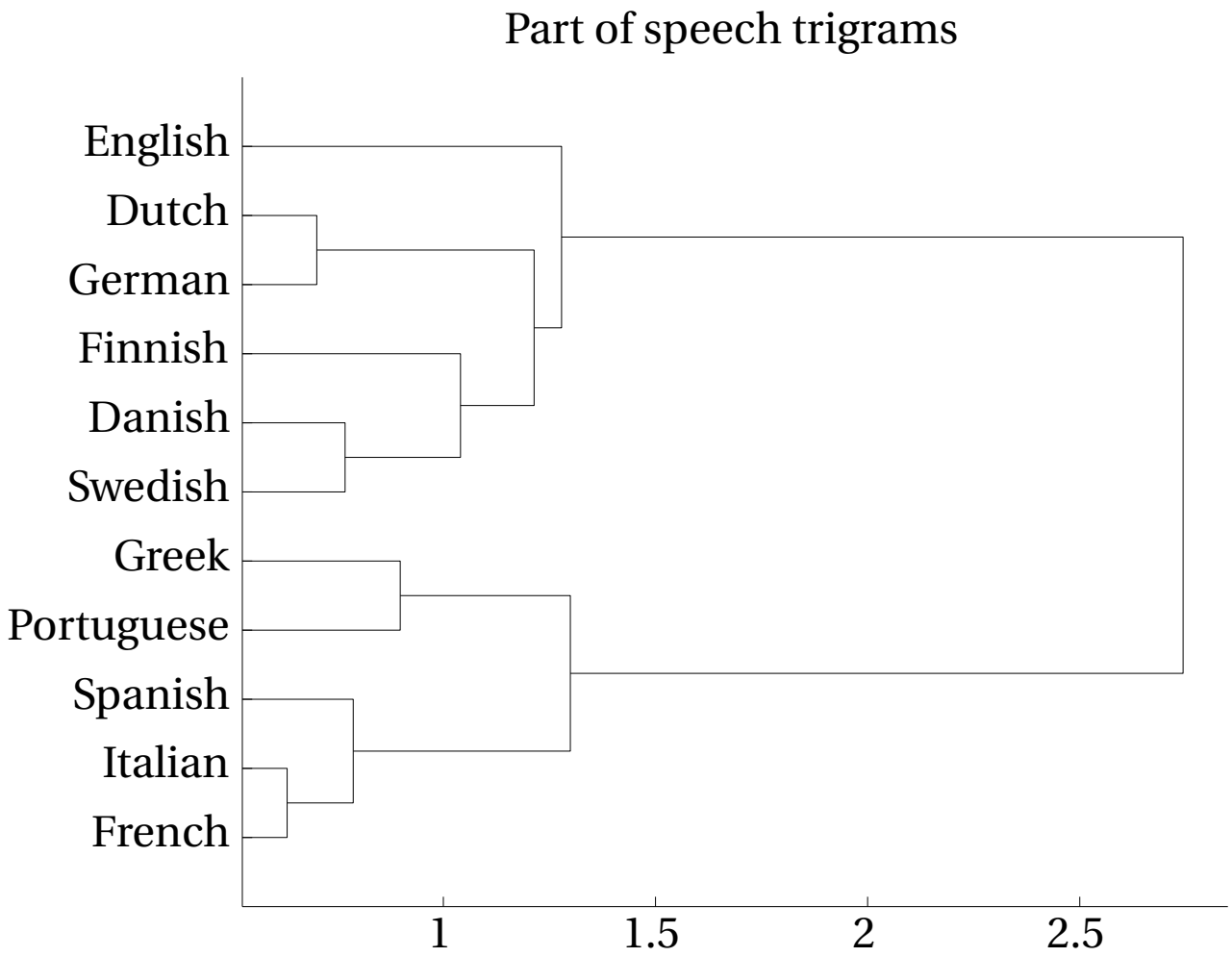
# Clustering of Translations

## Example (Clustering based on function words)



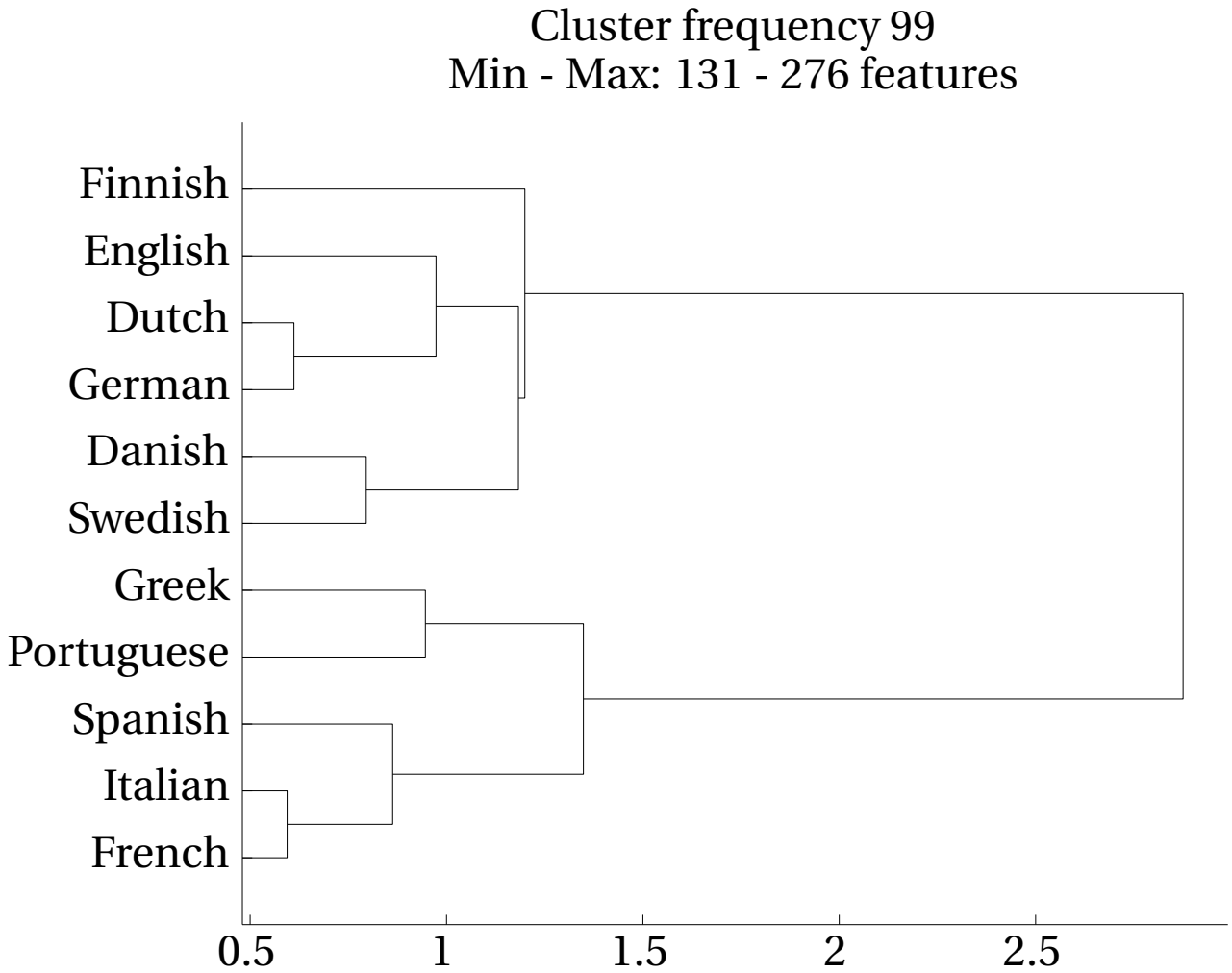
# Clustering of Translations

## Example (Clustering based on POS trigrams)



# Clustering of Translations

## Example (Frequent letter trigrams)



# Native Language Identification

- Given a set of essays composed by ESL students, identify the authors' native language (out of 11 languages)
- The dataset consists of TOEFL essays, contributed by the ETS
- Each text is annotated by the prompt, the proficiency level and the native language
- The task is to identify the native language
- Features
- Results (Tsvetkov et al., 2013)

# Native Language Identification

| | ARA | CHI | FRE | GER | HIN | ITA | JPN | KOR | SPA | TEL | TUR | P (%) | R (%) | $F_1$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ARA** | 80 | 0 | 2 | 1 | 3 | 4 | 1 | 0 | 4 | 2 | 3 | 80.8 | 80.0 | 80.4 |
| **CHI** | 3 | 80 | 0 | 1 | 1 | 0 | 6 | 7 | 1 | 0 | 1 | 88.9 | 80.0 | 84.2 |
| **FRE** | 2 | 2 | 81 | 5 | 1 | 2 | 1 | 0 | 3 | 0 | 3 | 86.2 | 81.0 | 83.5 |
| **GER** | 1 | 1 | 1 | 93 | 0 | 0 | 0 | 1 | 1 | 0 | 2 | 87.7 | 93.0 | 90.3 |
| **HIN** | 2 | 0 | 0 | 1 | 77 | 1 | 0 | 1 | 5 | 9 | 4 | 74.8 | 77.0 | 75.9 |
| **ITA** | 2 | 0 | 3 | 1 | 1 | 87 | 1 | 0 | 3 | 0 | 2 | 82.1 | 87.0 | 84.5 |
| **JPN** | 2 | 1 | 1 | 2 | 0 | 1 | 87 | 5 | 0 | 0 | 1 | 78.4 | 87.0 | 82.5 |
| **KOR** | 1 | 5 | 2 | 0 | 1 | 0 | 9 | 81 | 1 | 0 | 0 | 80.2 | 81.0 | 80.6 |
| **SPA** | 2 | 0 | 2 | 0 | 1 | 8 | 2 | 1 | 78 | 1 | 5 | 77.2 | 78.0 | 77.6 |
| **TEL** | 0 | 1 | 0 | 0 | 18 | 1 | 2 | 1 | 1 | 73 | 3 | 85.9 | 73.0 | 78.9 |
| **TUR** | 4 | 0 | 2 | 2 | 0 | 2 | 2 | 4 | 4 | 0 | 80 | 76.9 | 80.0 | 78.4 |

# Outline

# Conclusion

- Machines can accurately identify translated texts
- Translation "universals" should be reconsidered. Not only are they dependent on genre and register, they also vary greatly across different pairs of languages
- The best performing features are those that attest to the 'fingerprints' of the source on the target
- Interference, a pair-specific phenomenon, dominates other manifestations of translationese
- Translationese features are overshadowed by more salient features of the text, including genre, register, domain, etc.

# Conclusion

- Machines can accurately identify translated texts
- Translation "universals" should be reconsidered. Not only are they dependent on genre and register, they also vary greatly across different pairs of languages
- The best performing features are those that attest to the 'fingerprints' of the source on the target
- Interference, a pair-specific phenomenon, dominates other manifestations of translationese
- Translationese features are overshadowed by more salient features of the text, including genre, register, domain, etc.

# Conclusion

- Machines can accurately identify translated texts
- Translation "universals" should be reconsidered. Not only are they dependent on genre and register, they also vary greatly across different pairs of languages
- The best performing features are those that attest to the 'fingerprints' of the source on the target
- Interference, a pair-specific phenomenon, dominates other manifestations of translationese
- Translationese features are overshadowed by more salient features of the text, including genre, register, domain, etc.

# Conclusion

- Machines can accurately identify translated texts
- Translation "universals" should be reconsidered. Not only are they dependent on genre and register, they also vary greatly across different pairs of languages
- The best performing features are those that attest to the 'fingerprints' of the source on the target
- Interference, a pair-specific phenomenon, dominates other manifestations of translationese
- Translationese features are overshadowed by more salient features of the text, including genre, register, domain, etc.

# Conclusion

- Machines can accurately identify translated texts
- Translation "universals" should be reconsidered. Not only are they dependent on genre and register, they also vary greatly across different pairs of languages
- The best performing features are those that attest to the 'fingerprints' of the source on the target
- Interference, a pair-specific phenomenon, dominates other manifestations of translationese
- Translationese features are overshadowed by more salient features of the text, including genre, register, domain, etc.

# Conclusion

- Language models compiled from translated texts are better for SMT than ones compiled from original texts
- Translation models translated in the same direction as that of the SMT task are better than ones translated in the reverse direction
- Translation models can be adapted to translationese, thereby improving the quality of SMT

# Conclusion

- Language models compiled from translated texts are better for SMT than ones compiled from original texts
- Translation models translated in the same direction as that of the SMT task are better than ones translated in the reverse direction
- Translation models can be adapted to translationese, thereby improving the quality of SMT

# Conclusion

- Language models compiled from translated texts are better for SMT than ones compiled from original texts
- Translation models translated in the same direction as that of the SMT task are better than ones translated in the reverse direction
- Translation models can be adapted to translationese, thereby improving the quality of SMT

# Future Directions

- The features of Hebrew translationese: morphological markers (Avner et al., 2016)
- The features of **machine** translationese
- Robust identification of machine translation output
- More generally, the similarities and differences among various interference phenomena:

# Future Directions

- The features of Hebrew translationese: morphological markers (Avner et al., 2016)
- The features of **machine** translationese
- Robust identification of machine translation output
- More generally, the similarities and differences among various interference phenomena:
  - Translation
  - Non-native speakers
  - Learners' productions

# Future Directions

- The features of Hebrew translationese: morphological markers (Avner et al., 2016)
- The features of **machine** translationese
- Robust identification of machine translation output
- More generally, the similarities and differences among various interference phenomena:
  - Translation
  - Non-native speakers
  - Learners' productions

# Future Directions

- The features of Hebrew translationese: morphological markers (Avner et al., 2016)
- The features of **machine** translationese
- Robust identification of machine translation output
- More generally, the similarities and differences among various interference phenomena:
  - Translation
  - Non-native speakers
  - Learners' productions

# Future Directions

- The features of Hebrew translationese: morphological markers (Avner et al., 2016)
- The features of **machine** translationese
- Robust identification of machine translation output
- More generally, the similarities and differences among various interference phenomena:
  - Translation
  - Non-native speakers
  - Learners' productions

# Future Directions

- The features of Hebrew translationese: morphological markers (Avner et al., 2016)
- The features of **machine** translationese
- Robust identification of machine translation output
- More generally, the similarities and differences among various interference phenomena:
  - Translation
  - Non-native speakers
  - Learners' productions

# Future Directions

- The features of Hebrew translationese: morphological markers (Avner et al., 2016)
- The features of **machine** translationese
- Robust identification of machine translation output
- More generally, the similarities and differences among various interference phenomena:
  - Translation
  - Non-native speakers
  - Learners' productions

# Acknowledgements


- Gennadi Lembersky, Noam Ordan, Ella Rabinovich, Vered Volansky
- Israel Ministry of Science and Technology


# Publications


- Ella Rabinovich and Shuly Wintner. **Unsupervised Identification of Translationese**. *Transactions of the Association for Computational Linguistics* 3:419-432, 2015
- Vered Volansky, Noam Ordan and Shuly Wintner. **On the features of translationese**. *Digital Scholarship in the Humanities* 30(1):98-118, April 2015
- Gennadi Lembersky, Noam Ordan and Shuly Wintner. **Improving Statistical Machine Translation by Adapting Translation Models to Translationese**. *Computational Linguistics* 39(4):999-1023, December 2013
- Gennadi Lembersky, Noam Ordan and Shuly Wintner. **Language Models for Machine Translation: Original vs. Translated Texts**. *Computational Linguistics* 38(4):799-825, December 2012

# Bibliography I


Roee Aharoni, Moshe Koppel, and Yoav Goldberg. Automatic detection of machine translated text and translation quality estimation. In **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**, pages 289–295, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-2048`.

Omar S. Al-Shabab. **Interpretation and the language of translation: creativity and conventions in translation**. Janus, Edinburgh, 1996.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. Identifying translationese at the word and sub-word level. **Digital Scholarship in the Humanities**, 31(1):30–54, April 2016. doi: http://dx.doi.org/10.1093/llc/fqu047. URL `http://dx.doi.org/10.1093/llc/fqu047`.

Mona Baker. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, **Text and technology: in honour of John Sinclair**, pages 233–252. John Benjamins, Amsterdam, 1993.

Marco Baroni and Silvia Bernardini. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. **Literary and Linguistic Computing**, 21(3):259–274, September 2006. URL `http://llc.oxfordjournals.org/cgi/content/short/21/3/259?rss=1`.

Nitza Ben-Ari. The ambivalent case of repetitions in literary translation. Avoiding repetitions: A "universal" of translation? **Meta**, 43(1):68–78, 1998.

Shoshana Blum-Kulka. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, **Interlingual and intercultural communication Discourse and cognition in translation and second language acquisition studies**, volume 35, pages 17–35. Gunter Narr Verlag, 1986.

Shoshana Blum-Kulka and Eddie A. Levenston. Universals of lexical simplification. **Language Learning**, 28(2):399–416, December 1978.

Shoshana Blum-Kulka and Eddie A. Levenston. Universals of lexical simplification. In Claus Faerch and Gabriele Kasper, editors, **Strategies in Interlanguage Communication**, pages 119–139. Longman, 1983.

# Bibliography II


Andrew Chesterman. Beyond the particular. In Anna Mauranen and Pekka Kujamäki, editors, **Translation universals: Do they exist?**, pages 33–50. John Benjamins, 2004.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. **Computational Linguistics**, 16(1):22–29, 1990. ISSN 0891-2017.

Sauleh Eetemadi and Kristina Toutanova. Asymmetric features of human generated translation. In **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pages 159–164. Association for Computational Linguistics, October 2014. URL http://www.aclweb.org/anthology/D14-1018.

Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, **Translation Studies in Scandinavia**, pages 88–95. CWK Gleerup, Lund, 1986.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. **SIGKDD Explorations**, 11(1):10–18, 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL http://dx.doi.org/10.1145/1656274.1656278.

Qiwei He, Bernard P. Veldkamp, and Theo de Vries. Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. **Psychiatry research**, 198(3):441–447, 08 2012. URL http://linkinghub.elsevier.com/retrieve/pii/S0165178112000625?showall=true.

Iustina Ilisei. **A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models**. PhD thesis, University of Wolverhampton, Wolverhampton, UK, February 2013. URL http://clg.wlv.ac.uk/papers/ilisei-thesis.pdf.

Iustina Ilisei and Diana Inkpen. Translationese traits in Romanian newspapers: A machine learning approach. **International Journal of Computational Linguistics and Applications**, 2(1-2), 2011.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, **Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing**, volume 6008 of **Lecture Notes in Computer Science**, pages 503–511. Springer, 2010. ISBN 978-3-642-12115-9. URL http://dx.doi.org/10.1007/978-3-642-12116-6.

# Bibliography III


Patrick Juola. Authorship attribution. **Foundations and Trends in Information Retrieval**, 1(3):233–334, 2006. URL http:/dx.doi.org/10.1561/1500000005.

Dorothy Kenny. **Lexis and creativity in translation: a corpus-based study**. St. Jerome, 2001. ISBN 9781900650397.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In **Proceedings of the tenth Machine Translation Summit**, pages 79–86. AAMT, 2005. URL http://mt-archive.info/MTS-2005-Koehn.pdf.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions**, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-2045.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pages 1318–1326, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1132.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. **Literary and Linguistic Computing**, 14(3), 2003.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In **Proceedings of MT-Summit XII**, pages 81–88, 2009.

Sara Laviosa. Core patterns of lexical use in a comparable corpus of English lexical prose. **Meta**, 43(4):557–570, December 1998.

Sara Laviosa. **Corpus-based translation studies: theory, findings, applications**. Approaches to translation studies. Rodopi, 2002. ISBN 9789042014879.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. In **Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing**, pages 363–374, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D11-1034.

# Bibliography IV


Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Adapting translation models to translationese improves SMT. In **Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics**, pages 255–265, Avignon, France, April 2012a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E12-1026.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. **Computational Linguistics**, 38(4):799–825, December 2012b. URL http://dx.doi.org/10.1162/COLI_a_00111.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Improving statistical machine translation by adapting translation models to translationese. **Computational Linguistics**, 39(4):999–1023, December 2013. URL http://dx.doi.org/10.1162/COLI_a_00159.

Jianhua Lin. Divergence measures based on the shannon entropy. **IEEE Transactions on Information Theory**, 37(1): 145–151, January 1991. ISSN 0018-9448. doi: 10.1109/18.61115. URL http://dx.doi.org/10.1109/18.61115.

Lin Øverås. In search of the third code: An investigation of norms in literary translation. **Meta**, 43(4):557–570, 1998.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In **ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135.

Marius Popescu. Studying translationese at the character level. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, **Proceedings of RANLP-2011**, pages 634–639, 2011.

Rico Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In **Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics**, pages 539–549, Avignon, France, April 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E12-1055.

Sanjun Sun and Gregory M. Shreve. Reconfiguring translation studies. Unpublished manuscript, 2013. URL http://sanjun.org/ReconfiguringTS.html.

# Bibliography V


Elke Teich. **Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts**. Mouton de Gruyter, 2003.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. A report on the first native language identification shared task. In **Proceedings of the Eighth Workshop on Building Educational Applications Using NLP**. Association for Computational Linguistics, June 2013.

Gideon Toury. Interlanguage and its manifestations in translation. **Meta**, 24(2):223–231, 1979.

Gideon Toury. **In Search of a Theory of Translation**. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv, 1980.

Gideon Toury. **Descriptive Translation Studies and beyond**. John Benjamins, Amsterdam / Philadelphia, 1995.

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. Identifying the L1 of non-native writers: the CMU-Haifa system. In **Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications**, pages 279–287. Association for Computational Linguistics, June 2013. URL `http://www.aclweb.org/anthology/W13-1736`.

Hans van Halteren. Source language markers in EUROPARL translations. In Donia Scott and Hans Uszkoreit, editors, **COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK**, pages 937–944, 2008. ISBN 978-1-905593-44-6. URL `http://www.aclweb.org/anthology/C08-1118`.

Ria Vanderauwerea. **Dutch novels translated into English: the transformation of a 'minority' literature**. Rodopi, Amsterdam, 1985.

Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. **Digital Scholarship in the Humanities**, 30(1):98–118, April 2015.