The Soap Project

Bert Robberechts (Linguist)
Stefan Richter (Computer scientist)

Question

Question

Computer or Rechner?

What is the state of Anglicisms in German?

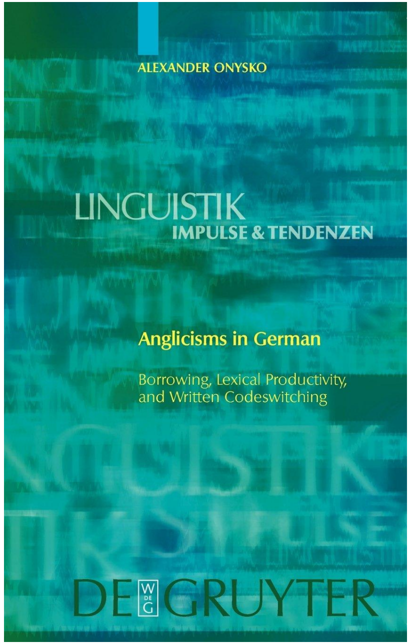
Which kind of words?

How often are they used?

When?

Where?

By whom?



ALEXANDER ONYSKO

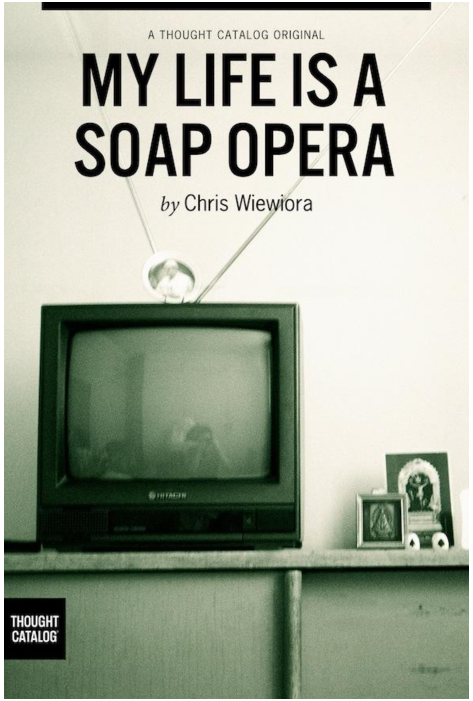
LINGUISTIK
IMPULSE & TENDENZEN

Anglicisms in German

Borrowing, Lexical Productivity, and Written Codeswitching

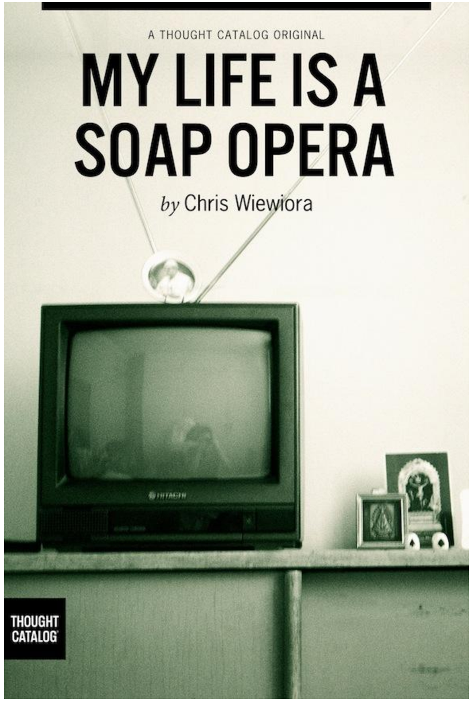
DE GRUYTER

Soap operas = a reflexion of life?



Soap operas = a reflexion of life?

How are Anglicisms used & distributed in German soap operas?



Possible uses

- Language teaching
- A historical perspective
- Contrastive: insight into sociolects, regional variety

Possible uses

- Language teaching
- A historical perspective
- Contrastive: insight into sociolects, regional variety

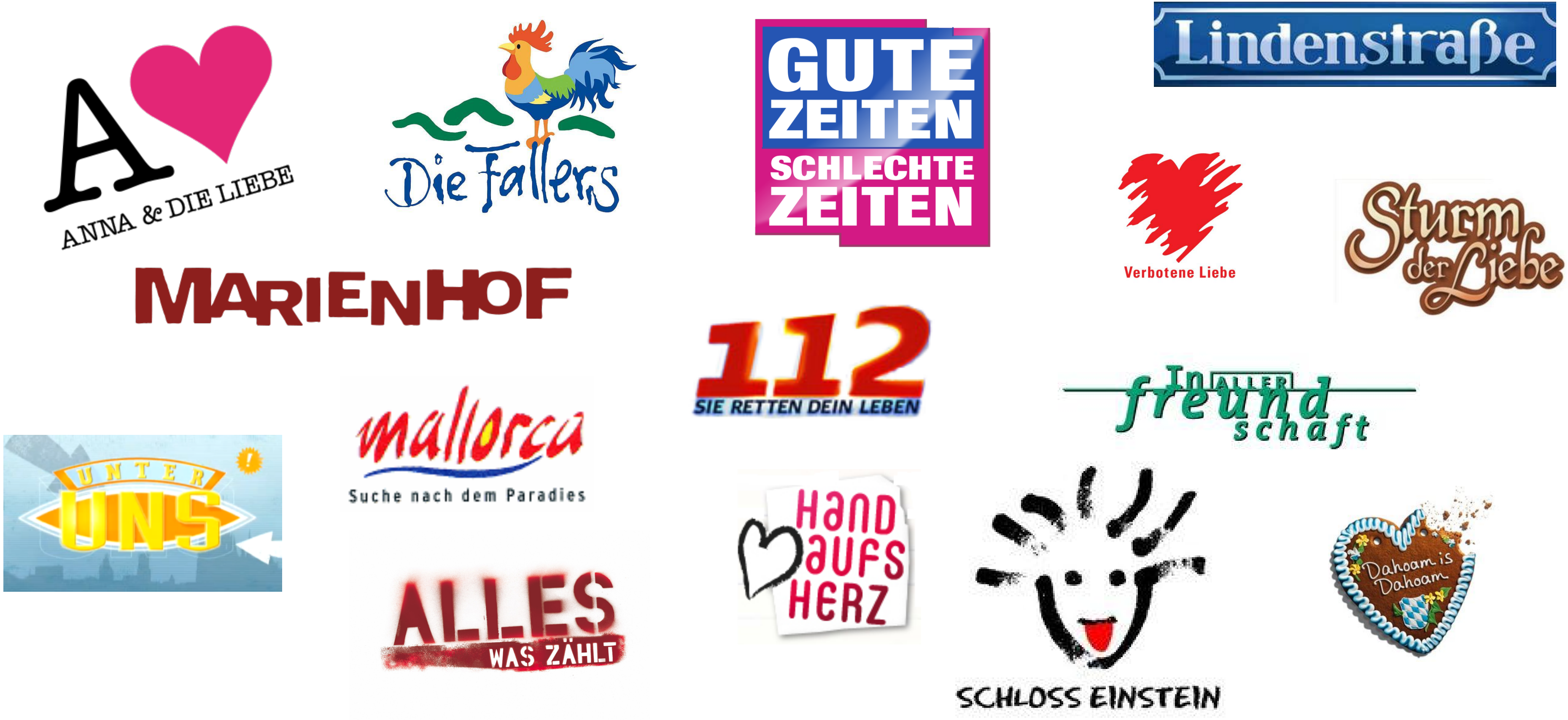
Possible uses

- Language teaching
- A historical perspective
- Contrastive: insight into sociolects, regional variety

Corpus

Start from scratch

German soap operas



German soap operas - subtitled

Lindenstraße

Verbotene Liebe

Sturm der Liebe

In aller freundschaft

Dahoam is Dahoam

German soap operas - subtitled

- ARD/ZDF (Public broadcasting)
- Online in Mediathek
  - video.flv + subtitle.xml

German soap operas - subtitled

- ARD/ZDF (Public broadcasting)
- Online in Mediathek
  - video.flv + subtitle.xml

- § 11d RStV
- Only available for 7 days

Lindenstraße

https://www.youtube.com/watch?v=aWxvXIh1mRU

Lindenstraße

- Weekly, since 1985
- Shows city life

- Subtitles were not removed from Mediathek
  - Crawled 193 episodes
  - 2010 - now
- *“A historical perspective”* ?

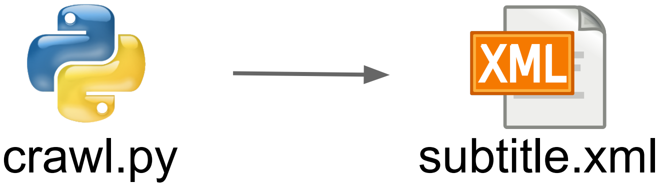
Dahoam is Dahoam

https://www.youtube.com/watch?v=KVSoAFPmlII

- Daily, since 2007
- Shows country life
  - Dialect

- Started crawling in April
  - Crawled 12 episodes
- *“Contrastive: sociolects + regional variety”* ?

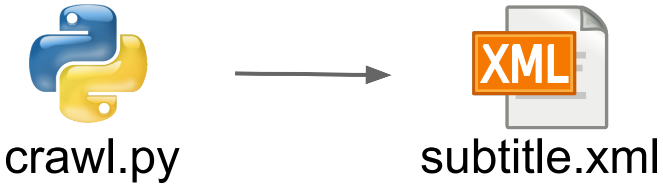
How we made our corpus



crawl.py

subtitle.xml

How we made our corpus

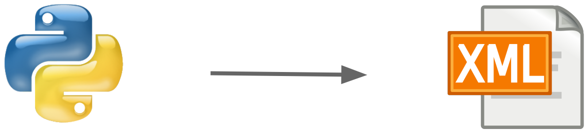


crawl.py

subtitle.xml

```
<p begin="366.8" end="368.6" tts:textAlign="center"> Vielleicht ein bisschen?  </p>
<p begin="369.2" end="371.0" tts:textAlign="center" tts:color="#9999FF"> Handy klingelt  </p>
<p begin="372.2" end="373.4" tts:textAlign="center"> Das ist Iffi.  </p>
<p begin="375.8" end="378.0" tts:textAlign="center"> Sie kommt über Weihnachten vorbei.  </p>
<p begin="380.0" end="382.5" tts:textAlign="center"> Wie gehts ihr?  <br />- Keine Ahnung.  </p>
<p begin="388.4" end="389.6" tts:textAlign="center"> Scheidung?  </p>
```

How we made our corpus



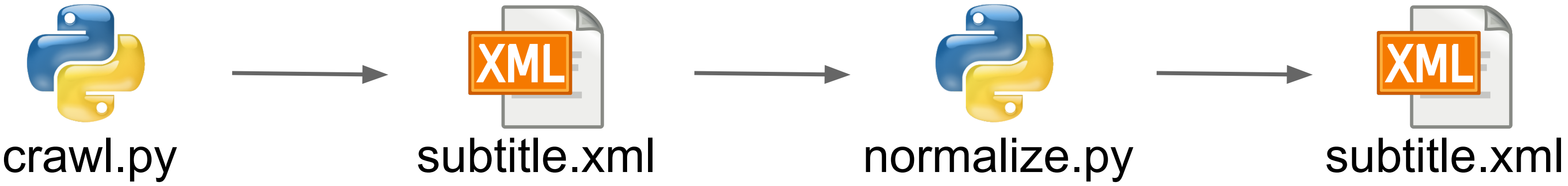
crawl.py

subtitle.xml

```
<tt:p xml:id="sub204" style="textCenter" region="bottom" begin="00:13:29.320" end="00:13:31.360">
        <tt:span style="textWhite">(flüstert) Hubert.</tt:span>
</tt:p>

<tt:p xml:id="sub205" style="textCenter" region="bottom" begin="00:13:31.440" end="00:13:35.360">
        <tt:span style="textGreen">Ja, was ist denn?</tt:span>
        <tt:br />
        <tt:span style="textWhite">Die Franzi hat keinen Schnuller.</tt:span>
</tt:p>
```

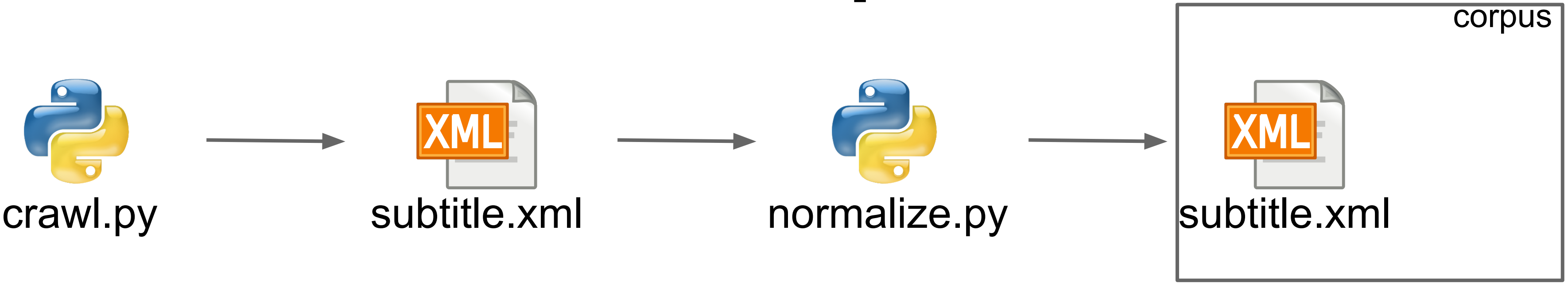
How we made our corpus



crawl.py → subtitle.xml → normalize.py → subtitle.xml

```
<xml episode="1286" name="Lindenstrasse" year="2010">
        <p begin="19.2" end="21.6">Iffi, lass ihn in Frieden. </p>
        <p begin="22.0" end="24.6">Er zieht sich so sehr zurück. </p>
        <p begin="25.1" end="27.9">Hey, lasst das. Er verzweifelt. </p>
        <p begin="28.3" end="32.9">Er ist zu alt, um sich von </p>
        <p begin="46.9" end="50.7">Hey, weißt du, wie spät es ist? </p>
        <p begin="51.1" end="53.1">Was schaust du da? </p>
```

How we made our corpus



crawl.py

subtitle.xml

normalize.py

corpus

subtitle.xml

- Corpus 340.395 words
  - Lindenstrasse: 304.729 words in 193 episodes
  - Dahoam is Dahoam: 35.666 words in 12 episodes
- ©ARD

Our process & methods

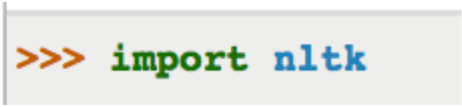
Chop it up

Check if the tokens are in an English wordlist

Our process & methods

Chop it up

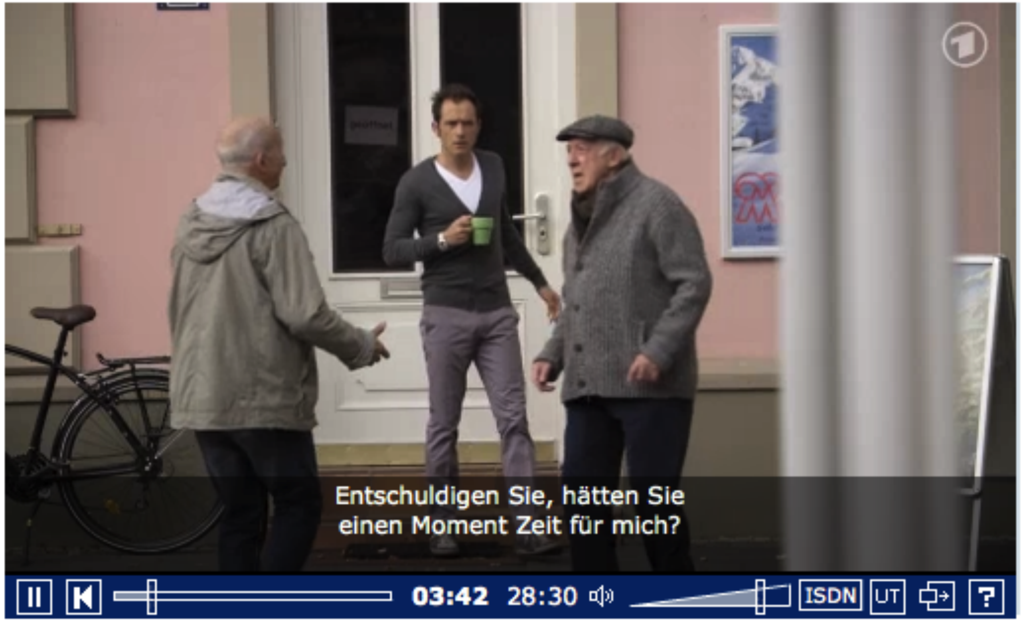
Check if the tokens are in an English wordlist



```
>>> import nltk
```

Using the **Natural Language Toolkit**

Folge 1459 "Oktober suchen", 08.12.2013



Entschuldigen Sie, hätten Sie
einen Moment Zeit für mich?

1304.xml: <p begin="364.2" end="366.7">Hast du einen *Moment* für mich? </p>

It’s raining noise

It’s raining noise

```
1304.xml:  <p begin="364.2" end="366.7">Hast du einen Moment für mich? </p>
1356.xml:  <p begin="810.0" end="812.8">Er kommt dann mit "Pilot" </p>
```

Is it English | Ist es Deutsch?

Improving the algorithm to deal with ambiguity

The question remains: when is a borrowed word fully part of a language?
Where will we draw the line?
How can we improve our results?

A: Part of Speech Tagging

B: Colocation

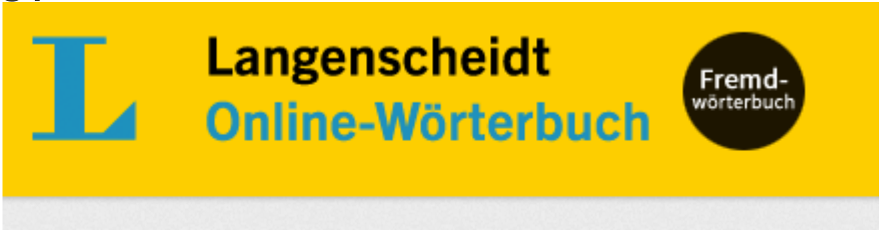
C: Cross-referencing with a German Wordlist

D: Scrape information from online dictionaries

Is it English | Ist es Deutsch?

Improving the algorithm to deal with ambiguity

The question remains: when is a borrowed word fully part of a language?
Where will we draw the line?
How can we improve our results?



A: Part of Speech Tagging

B: Colocation

C: Cross-referencing with a German Wordlist

D: Scrape information from online dictionaries

Is it English | Ist es Deutsch?

Improving the algorithm to deal with ambiguity

The question remains: when is a borrowed word fully part of a language?
Where will we draw the line?
How can we improve our results?

A: Part of Speech Tagging

B: Colocation

C: Cross-referencing with a German Wordlist

D: Scrape information from online dictionaries

E: A noise list

Is it English | Ist es Deutsch?

Improving the **y**

The question re ully part of a language?
Where will we
How can we im

A: Part of

B: Colocat

C: Cross-r :

D: Scrape information from online dictionaries

E: A noise list

Analysis

- Statistical analysis
    - **Frequency**
    - Change of **usage** over **time**
    - Relation between **user groups** and **usage**
    - Collocation: where do Anglicisms appear?

Analysis

- Qualitative analysis
    - **Semantical categories**

Outlook

What we have

- Corpus
  - Lindenstrasse (193 episodes)
  - Dahoam is Dahoam (12 episodes)
- Simple English language detection algorithm
  - based on nltk
- 8 weeks of “Vorlesungszeit”

What we need to do

- Improve English language detection algorithm
- Quantitative analysis
- Qualitative analysis
- Visualization of results
- Detection of loanwords
    - gedownloaded
    - gecherrypicked

When do we do it

	CW 22	CW 23	CW 24	CW 25	CW 26	CW 27	CW 28	CW 29	CW 30+
Bert	Improve English language detection algorithm				Qualitative analysis				Detection of loanwords
Stefan					Quantitative analysis		Visualization of results		

When do we do it

	CW 22	CW 23	CW 24	CW 25	CW 26	CW 27	CW 28	CW 29	CW 30+
Bert	Improve English language detection algorithm				Qualitative analysis				Detection of loanwords
Stefan					Quantitative analysis		Visualization of results		

- If we finish early
  - Extend corpus
  - Improve algorithms

When do we do it

<table>
<tr><th></th><th>CW 22</th><th>CW 23</th><th>CW 24</th><th>CW 25</th><th>CW 26</th><th>CW 27</th><th>CW 28</th><th>CW 29</th><th>CW 30+</th></tr>
<tr><td><b>Bert</b></td><td colspan="4" rowspan="2">Improve English language detection algorithm</td><td colspan="4">Qualitative analysis</td><td rowspan="2">Detection of loanwords</td></tr>
<tr><td><b>Stefan</b></td><td colspan="2">Quantitative analysis</td><td colspan="2">Visualization of results</td></tr>
</table>

- If we finish early
  - Extend corpus
  - Improve algorithms
- If we have problems
  - No visualization
  - Reduce qualitative analysis

First results

...don’t trust them

Lindenstraße

$$\frac{\text{english words}}{\text{total words}} * 100 = 0{,}68\%$$

Dahoam is Dahoam

$$\frac{\text{english words}}{\text{total words}} * 100 = 0{,}53\%$$

Lindenstraße

$$\frac{\text{english words}}{\text{total words}} * 100 = 0{,}68\%$$

51	Job
39	Handy
30	Hi
25	Party
24	Society
22	Chef
18	Internet
14	Video
13	Computer
12	cool

$$\frac{\text{english words}}{\text{total words}} * 100 = 0{,}53\%$$

6	Handy
5	Tablet
5	Date
4	Mike
4	F
3	Designer
3	Gaudi
3	Baum
3	cut
3	Tumor

Sources of inspiration

Anglicisms in German and other languages

Anglicisms in German: Borrowing, Lexical Productivity, and Written Codeswitching.

Onysko, Alexander and Walter de Gruyter, 2007

The impact of nominal anglicisms on the morphology of modern spoken German.

Hunt, Jaime. 2011.

Investigating loan words and expressions in tourism discourse: a corpus driven analysis on the BBC-travel corpus.

Gandin, Stefania in European Scientific Journal, Jan 15, 2014, Vol.10(2)

Sources of inspiration

On corpus-based research into lexical borrowing

Pimp my Lexis: het nut van corpusonderzoek in normatief taaladvies

Van de Velde, Freek ; Zenner, Eline

A usage-based onomasiological approach to measuring the success of loanwords

Zenner, Eline

Tendencies of using loan words in the Corpus of the Lithuanian language Rudzevičius in Povilas Kalbos kultūra, 2005, Issue 78, p. 210–219

Corpora as lexicographical basis – The case of anglicisms in Norwegian

Gisle Andersen

Sources of inspiration

Non-academic publications on Anglicisms in German

http://www.brighthubeducation.com/learning-german/74147-anglicisms-in-german-language/

http://www.theguardian.com/world/2011/feb/01/german-language-english-words-leaken

On Anglicisms in general

The Anglicization of European Lexis

Ed. Cristiano Furiassi, Virginia Pulcini, Félix Rodríguez González

Discussion & Future