PROJECT REPORT

Improving DBpedia Spotlight for Dutch*

Semantic Web Technology

* and German and English

David de Boer
Joachim Daiber

Contents

1 INTRODUCTION

In this project, we will integrate and streamline the work that has been done in Google Summer of Code 2012, create an improved Dutch version of the DBpedia Spotlight semantic annotation system and write an Internationalization tutorial for the new system (while performing all necessary tasks to make the internationalization as simple as possible). We will evaluate the system on heldout data and compare it to the current Dutch version. Furthermore, we will improve the Spotting for Dutch and manually evaluate the quality of Spotting for different algorithms.

Our own criteria for the evaluation of this project are that the new system must be *faster* and *more accurate* than the version currently available for Dutch. Training the system must be *easier* and *faster* than the current system and the internationalization guide, which we provide as a deliverable, must be easy to understand and reproduce.

2 RELATED WORK AND CONTEXT

DBpedia Spotlight (Mendes et al., 2011) is an open source project developing a system for the automatic annotation of DBpedia entities in natural language text. In the following, we will refer to the currently available version of DBpedia Spotlight as the *current build*. The new version of the system will be referred to as *our build*. This version incorperates source code produced over the course of Google Summer of Code 2012 by Joachim Daiber and other students as well as new additions to the source code that were part of this project.

The current build of DBpedia Spotlight provides general interfaces for spotting (recognition of phrases to be annotated) and disambiguation as well as various output formats (XML, JSON, RDF, etc.) in a REST-based[1] web service. The disambiguation is based on a modifcation of TF-IDF (using Apache Lucene) and the main Spotting algorithm is an exact dictionary based lookup using LingPipe.[2] As LingPipe is under a non-free license, it is problematic to include it in the Apache 2.0 licensed Spotlight code.

Our build uses neither Apache Lucene nor LingPipe. In our build, the indexing process of a model consists of two parts:

- **Step 1: raw data collection with PigNLProc.** Apache Pig is a MapReduce-based platform for working with large data sets on Hadoop clusters. Pig scripts are written in an SQL-like syntax. PigNLProc is a collection of pig scripts and utilties focused on Wikipedia and DBpedia, which are used to produce all necessary counts for the model.

- **Step 2: Creating the model from the raw data.** The raw counts from step 1 are serialized into efficient data structures that the system can deserialize at runtime.

[1] Representational state transfer.
[2] Alias-i. 2013. LingPipe 4.1.0. `http://alias-i.com/lingpipe`

This step can be performed on a regular PC and takes a few minutes for Dutch and 1-2 hours for the English Wikipedia, which contains significantly more data than all other Wikipedia dumps.

3 OUR APPROACH

3.1 Indexing and models

3.1.1 Collecting occurrence counts

To select the correct annotation candidates from the set of candidates we generate in various ways (see Section 3.2), we estimate the probability that a surface form (sf) is annotated as:

$$\mathrm{P}(\mathsf{annotation}|\mathrm{sf}) = \frac{\sum_e \mathsf{count}(e, \mathrm{sf})}{\mathsf{count}(\mathrm{sf})}$$

The value of $\mathsf{count}(\mathrm{sf})$ is the total number of times the surface form occurs as a string in the whole dataset. Since it would not be feasible to perform a string search for each surface form (>5m in the case of English), in the current version of PigNLProc, ngrams (with n up to 5 by default) are extracted from the whole dataset and are then joined with the set of surface forms. This approach requires the temporary storage of all possible ngrams, which produces a noticeable bottleneck. Hence, we extended PigNLProc to only collect ngrams for the set of accepted surface forms, which is distributed to the cluster nodes via the Hadoop distributed cache.

When we analyzed the values of P("annotation"|sf) for Dutch, we observed that for some entities, the probabilities were consistently lower. After further investigation, we found that these were all cases where the surface form is a substring of another surface form. In the Pig script, $\mathsf{count}(e, \mathrm{sf})$ only considers full annotations and not their parts or tokens. Since $\mathsf{count}(\mathrm{sf})$ is the general frequency of a surface form, cases where a surface form is contained in another surface form are counted as a non-annotation of the contained surface form. Consider, for example, the surface form "Apple" and suppose that our corpus consists of the following text (where [...] indicates an annotation):

```
[Apple] is the company selling the [Apple MacBook], the [Apple iPod] and
the [Apple iPad].
```

According to the counts produced by the Pig script, in this corpus $\sum_e \mathsf{count}(e, \text{"Apple"})$ would be 1 and $\mathsf{count}(\text{"Apple"})$ would be 4. Hence $\mathrm{P}(\mathsf{annotation}|\text{"Apple"}) = \frac{1}{4}$. However, the annotated mentions of "Apple MacBook", "Apple iPod" and "Apple iPad" should not count as unannotated occurrences of the surface form "Apple" since there can be only one annotation in the text span "Apple X" and not both annotations (e.g. `[[Apple] iPad]`). Hence, in our build, we correct the counts from the Pig script

by subtracting the annotated counts of bigger surface forms from the total counts of their surface form substrings.

Furthermore, we added the ability to separate a set of heldout data from the dump that we use for tuning the spotter and the disambiguator and that can be used for testing.

3.1.2 Models and configuration

In the current build, a model for a specific language consists of a configuration file and various files that are spread over various directories. To avoid unnecessary complexity, we introduced a self-contained model directory structure that is produced by the indexing part and can be ran using the Spotlight server (see the Internationalization guide for more details).

3.2 Spotting

For the spotting part, we leverage existing methods for detecting noun phrases and named entities and we resolve overlaps and remove false positives according with our own model. Apache OpenNLP supplies pre-built models for chunking and named entity recognition which can be integrated in the DBpedia Spotlight code. We are using the Dutch named entity regonizer provided by OpenNLP. Since no noun phrase chunking models were available for Dutch, we trained this model using the OpenNLP toolkit.

3.2.1 OpenNLP-based spotter

Some of the data required for the spotting task are already available within the system as part of the disambiguation task, hence one of our goals was to minimize the addional memory and disk footprint of the spotting task. The interface `SurfaceFromStorage` is used in the disambiguation to retrieve additional information for a surface form candidate (its unique ID and its annotation probability).

We developed a Spotter that proceeds in two steps. In the first step, all spot candidates are generated using three methods:

1. Identify all sequences of capitalized tokens.
2. Identify all named entities.
3. Identify all noun phrases, prepositional phrases and multi word units.

Method 2 and 3 are performed using Apache OpenNLP models. The Spotter remains functional if no models are available (in this case, it only uses method 1), however for the best performance at least the chunker model should be included.

In the second step, each spot candidate is assigned a score and overlaps in the candidates are resloved using their score and the method that generated them. We

are using the following precedence relation (from most preferred to least preferred): PER $\succ$ ORG $\succ$ LOC $\succ$ MISC $\succ$ NP $\succ$ MWU $\succ$ PP $\succ$ Capitalized Sequence. All candiates that fall below a specified score threshold are removed.

As the score for a candidate, we initially used only P(annotation | sf) as defined above. However, in the resulting annotations, we identified several cases where the annotation probability for a string is consistently lower than in the general case. Examples include acronyms (e.g. 'AIG', 'IBM') and surface forms that are substrings of other surface forms (e.g. 'Bush' of 'George W. Bush'). To be able to treat such cases differently and be more flexible, we compute the score as a linear combination of scores, currently: P("annotation" | sf), a value indicating if a surface form is an acronym and the bias value 1.0. We estimate the weights from heldout data via linear regression. This method provides the additional advantage that it automatically estimates the optimal annotation threshhold for the data.

Because we use the methods above for generating phrase candidates, we can simply use the existing `SurfaceFromStorage` interface to calculate the score for annotation candidates and hence, apart from the OpenNLP models, we do not increase the memory usage.

3.2.2 Dutch chunking

Since Apache OpenNLP doesn't supply a phrase chunker model for Dutch, we performed the training ourselves. The OpenNLP toolkit contains a training tool for creating a chunker model, which requires a dataset of annotated sentences.

For the creation of this dataset, we used the LASSY Small corpus, which is available for RuG students on the university's network. LASSY Small is a corpus of Dutch language texts, automatically syntactically annotated and manually corrected (van Noord et al., 2009).

Using XQuery, we could extract chunks from the corpus to create a dataset of Noun Phrase chunks (NP), Prepositional Phrase chunks (PP) and Multi Word Units (MWUs) from the LASSY Small Corpus. We chose to add PP-chunks and MWUs since these phrases are likely to contain the entities we want to semantically annotate.

The table below show the results of the OpenNLP toolkit's cross validatior for the model we created using the OpenNLP toolkit.

	Precision	**Recall**	**F1**
PP	90.41	93.03	91.7
NP	83.19	80.75	81.95
MWU	87.27	75.1	80.73
Total	86.1	84.04	85.06

3.3 Disambiguation

Disambiguation in our build is performed using the generative probabilistic model from (Han and Sun, 2011) that was implemented by Joachim Daiber before the start of this project. The score for each entity e given the surface form s and context c is calculated as a combination of $\mathrm{P}(e)$, $\mathrm{P}(s|e)$ and $\mathrm{P}(c|e)$. The combination can be either a linear regression mixture or the product of the values.

In this model, probabilities are negligibly small and are represented in logarithmic space to avoid underflows. To produce a more useful score for each entity, we normalize the score via the commonly used sigmoid function and obtain a final disambiguation score between 0.0 and 1.0.

We extended this implementation by adding handling of `NIL` entities. For each surface form and context, we generate a `NIL` entity, which represents the null hypothesis that the context and surface form were not generated by any known entity but were random. We use the formulation from Han and Sun (2011) to calculate the score for the `NIL` entity:

$$\mathrm{P}(\texttt{NIL}) = \frac{1}{|M|} \tag{1}$$

$$\mathrm{P}(s|\texttt{NIL}) = \prod_{t \in S} \mathrm{P}_{\mathrm{LM}}(t) \tag{2}$$

$$\mathrm{P}(c|\texttt{NIL}) = \prod_{t \in C} \mathrm{P}_{\mathrm{LM}}(t) \tag{3}$$

In this case, P_{LM} is the smoothed general language model probability of a token that we estimate over all tokens imported to the system as context of an entity mention. The individual scores are combined as for all other entities and all entities with a lower score than the `NIL` entity are removed.

4 EVALUATION

Reiterating the main evaluation criteria set out in the introduction, our aims were to provide (1) easier and faster indexing, (2) faster and more accurate runtime performance and (3) simple internationalization. In order to demonstrate criteria (1) and (3), in addition to the Dutch model, we also created models for English and German. Demonstrations of the systems are available at:[3]

- Dutch: `http://jodaiber.github.com/demo/index.html`
- German: `http://jodaiber.github.com/demo/index_de.html`
- English: `http://jodaiber.github.com/demo/index_en.html`

[3] All systems run on a standard personal computer with 24GB of memory. If they should be unavailable, please contact the authors.

For details on criteria (3), i.e. the indexing process and internationalization, please see our internationalization guide in the GitHub project wiki.[4]

Evaluation was carried out on both our build as well as the current build of DBpedia Spotlight for Dutch. We evaluated indexing and runtime performance, spotting and disambiguation. This way we can compare the overall performance of both systems.

4.1 Performance: indexing and runtime

4.1.1 Model size

Language	Model	Disk space requirement
Dutch	Our build	489MB
	Current build	2.1GB
German	Our build	1.4GB
	Current build	-
English	Our build	5.2GB
	Current build	18.2GB

Since the models in the current version of our build are fully loaded into memory, they require less disk space. In the current build, on the other hand, the system relies mostly on disk access via Apache Lucene. These models are optimized for efficient disk-based access and therefore require more disk space.

4.1.2 Memory footprint

Language	Model	Memory usage
Dutch	Our build	1.9GB
	Current build (1)	2.41GB
	Current build (2)	14.6GB
German	Our build	4.8GB
	Current build	-
English	Our build	11.7GB
	Current build (1)	5.96GB

The table above shows the memory usage of each of the systems. All systems were run on the same personal computer, a 6 core AMD Phenom x6 with 24GB of memory

[4] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Internationalization-(DB-backed-core)`

and an SSD drive. We measured the memory usage after the system annotated a single text immediatelly after startup. In current build (1) the candidate index is kept in memory and and the disambiguation index is on disk and in current build (2) both indexes are loaded to memory. Since in the version of our build all data is kept in memory (contextual data can also be kept on disk), our English model requires more memory than the current build. However, the English model of the current build is still large since the LingPipe-based exact dictionary Spotter builds a trie data structure from all possible surface forms, which is a memory-intensive process. In our build, we can avoid this by using the data already in memory with our OpenNLP Spotter implementation.

4.1.3 Annotation time

To test the annotation time performance, we annotated a small corpus of around 500 articles for each language using all systems with their default settings. Except for the current build English version marked as *Public endpoint*, all tests were performed on the same personal computer.

Language	Model	Avg time/article	Total time
Dutch	Our build	1.20s	601.39s
	Current build (1)	9.52s	4758.31s
German	Our build	1.07s	504.68s
	Current build	-	-
English	Our build	1.28s	640.2s
	Public endpoint	5.72s	2803.22s

Additional to our own tests, a company that was interested in our build ran performance tests on a dedicated computing instance on the Amazon Web Services platform. Both systems were run on an instance of type `m1.xlarge`, which has 4 CPU cores and 15GB of memory. We show here the response times for our English model and an English model using an older version of Spotlight (0.6.5), which was faster than the current version. The full results are available online.[5]

Model	Avg time/article	Min time/article	Max time/article
Our build (1)	0.613s	0.017s	10.751s
Our build (2)	0.398s	0.010s	16.127s
Spotlight 0.6.5	0.223s	0.011s	26.8s

(1) Spotting with POS tagging and chunking,
(2) Spotting without POS tagging and chunking

[5] `http://faveeo.github.com/faveeo/perftests/`

4.2 Spotting

Spotting was evaluated manually. For this we selected nine texts on different topics: news articles, movie reviews and tourist information.

We used news articles from the *volkskrant97* corpus, which is available on the RuG network. The movie reviews were obtained from `http://nu.nl/filmrecensies` and the tourist information was obtained from `http://reistipseuropa.nl`.

Nine texts were automatically annotated using our build (available on `http://jodaiber.github.com/demo`) and the current build of DBpedia Spotlight for Dutch, which is available online at `http://nl.dbpedia.org/spotlight`.

To evaluate the spotting performance, a manual count was performed of the number of correctly annotated entities, incorrectly annotated entities and the entities that were not annotated at all. From these numbers we can derive precision and recall using the following formulas:

$$Precision = \frac{tp}{tp + fp}$$
$$Recall = \frac{tp}{tp + fn}$$

To give a better indication of the overall performance, the F1-measure is calculated from the precision and recall values using the formula below.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

The results are displayed in the table below. The exact counts and used texts are available in Appendix A.

Model	**Precision**	**Recall**	**F1-measure**
Our build (OpenNLP spotter)	92.74	46.13	61.61
Current build (LingPipe spotter)	10.34	35.67	16.03

The comparison shows a significant increase in precision, recall and the F1-score for our build of DBpedia Spotlight for Dutch. The greatest increase is seen in the precision and the F1-measure of our build. The high precision and moderate recall show that the OpenNLP spotter acts rather conservative in which entities it annotates, hence it might be necessary, depending on the task at hand, to lower the annotation threshold which will decrease the precesion but increase the recall.

To check these results, we also performed an automatic evaluation of the spotting performance on heldout data from Wikipedia. We used the first 10.000 paragraphs of the same heldout data we also use for testing the disambiguation performance. Note that the results are bound to show lower scores than the manual results since the annotation style in Wikipedia differs from our target annotation style, e.g. in contrast to our OpenNLP spotter, in Wikipedia previously mentioned entities are not annotated while verbs and adjectives are often annotated.

Model	**Precision**	**Recall**	**F1-measure**
Our build (OpenNLP spotter)	49.45	55.53	52.32
Current build (LingPipe spotter)	8.26	77.17	14.92

4.3 Disambiguation

We evaluated the disambiguation performance for Dutch of both builds automatically on heldout data from Wikipedia. Disambiguation was performed on randomly selected heldout data paragraphs, which were filtered to only include annotations that are ambiguous (the surface form has more than one possible corresponding entity) and that are not disambiguation pages. After this filter step, 28.475 paragraphs remain. The evaluation assumes that a surface form in a text is given and the task is only to find the correct entity for the surface form. *Accuracy* is the percentage of correctly predicted entities *MRR* is the mean reciprocal rank, indicating the accuracy of the n-best predictions and *URI not found* is the percentage of cases in which the gold entitiy was not among the n-best predictions.

Note that the current build for Dutch is trained on the full Wikipedia data, including the heldout sections we use for evaluation. We show results of our build both using a model trained on only the training section of the Wikipedia dump (train) and on the full dump including the heldout sections (full). Current build D1 is uses the `MixedWeightsDisambiguator` class, current build D2 uses `TwoStepDisambiguator`. We also reprint the results that were measured on an independent test corpus (Milne and Witten, 2008) for the English build at the end of Google Summer of Code 2012.

Language	**Model**	**Accuracy**	**MRR**	**URI not found**	**Time**
Dutch	Our build (train)	0.841	0.622	0.067	614s
	Our build (full)	0.883	0.646	0.055	613s
	Current build D1	0.581	0.432	0.367	6650s
	Current build D2	0.512	0.399	0.354	6004s
English	Our build	0.851	0.7978	0.074	-
	Current build D1	0.716	0.6883	0.166	-

5 DISCUSSION AND CONCLUSION

Our Dutch version is significantly faster than the current Dutch endpoint. To our surprise, we noticed that the English Spotlight version deployed to `http://spotlight.dbpedia.org` showed faster response times than our memory-based English version in some cases. However, as we have seen in Section 4.1.2, the price for a fast verion of the current build is an extreme memory footprint, significantly higher than the upper bound of our build. The tests run on the Amazon Web Service instance also showed that an old version of DBpedia Spotlight is faster than our build. This evaluation further indicated that the additional overhead introduced by the part-of-speech tagging and chunking in our OpenNLP spotter is only about 0.2s per article. Hence, there should still be room for improvement in the performance of the memory-based system, especially in the disambiguation algorithm.

While our linguistically-motivated spotting method (using phrase chunks and named entities) provides high accuracy, it is also slower than more generic algorithms like exact dictionary-based spotting. This can be observed in the time difference between our build and the current build in the disambiguation task and the overall annotation time as well as in the results from the Amazon Web Service instance. As future work, it might be beneficial to explore alternative methods of generating spot candidates that are faster than the OpenNLP-based candidate generation.

We are confident that we achieved the goals set out for this project. Our Dutch system shows significantly better accuracy, both in disambiguation and in spotting, the annotation time was greatly reduced compared to the current version and is close to the performance of earlier versions of DBpedia Spotlight. We were able to achieve these results by extending and streamlining the work that was done in Google Summer of Code 2012. Furthermore, we demostrated the ease of internationalization with our system by creating Dutch, German and English endpoints.

REFERENCES

[Han and Sun2011] Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 945--954. Association for Computational Linguistics.

[Mendes et al.2011] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *I-SEMANTICS*, pages 1--8.

[Milne and Witten2008] David N. Milne and Ian H. Witten. 2008. Learning to link with Wikipedia. In *CIKM*, pages 509--518.

[van Noord et al.2009] Gertjan van Noord, Gosse Bouma, Frank van Eynde, and Daniel de Kok et al. 2009. *Large Scale Syntactic Annotation of Written Dutch: Lassy*.

A MANUAL EVALUATION COUNTS

These texts were annotated automatically using DBpedia Spotlights builds on http://jodaiber.github.com/demo/ (our build) and http://nl.dbpedia.org/spotlight/ (Current Build) on default settings.

TEXT 1 - NEWS ARTICLE - Stijve Van Aartsen kent zijn dossiers - BINNENLAND - Volkskrant 971008

Annotated by our dutch model

DE MINISTER VAN VIS, vlees, groenten, zee en bos houdt niet van vlees. Hij eet het, maar als hij mag kiezen, neemt hij liever vis. Jozias van Aartsen is daarentegen dol op de natuur en trekt er graag op uit.
Soms neemt hij de fiets om van zijn huis naar het ministerie te gaan. Dat kan goed, want niemand herkent de minister. Zelfs toen de mestwet in de Tweede Kamer werd behandeld en boze boeren het departement belegerden, liet hij een keer de dienstauto staan en kwam hij ongedeerd binnen.

Het ministerschap van Van Aartsen is gedomineerd door de gekkekoeienziekte en de varkenspest. De bewindsman heeft die crises als een behoedzaam manager behandeld.
Hij kondigde bij de BSE direct ingrijpende maatregelen af. Het vlees verdween van de schappen. Bij de varkenspest reageerde de minister trager, voorzichtiger. Het duurde weken voor hij een fokverbod afkondigde. Zijn adviseurs waren woedend over die trage besluitvorming. Van Aartsen verdedigde zich: ingrijpende besluiten moeten zorgvuldig worden genomen. Toch verweet de Kamer hem dat hij achter de feiten aan rende.

De twee veeziekten die het land teisterden, beperkten Van Aartsen in zijn bewegingsvrijheid. Hij had nauwelijks tijd eigen beleid te ontwikkelen. De gekkekoeienziekte met de BSE-crisis hield hem in 1996 goeddeels van de straat, en dit jaar was het de varkenspest die het gewone regeren onmogelijk maakte.

Het Europese voorzitterschap kostte ook al veel tijd, maar leverde weinig tot de verbeelding sprekende successen op. Dat kan ook haast niet, want het Europese landbouwbeleid is een zaak van lange adem. Snelle succesjes zijn zeldzaam.

Met de mestwet had de minister wel kunnen scoren, maar daar liep het slecht mee af. Van Aartsen trotseerde het boerengeweld buiten de zaal, maar week uiteindelijk voor druk uit de Kamer. De wet werd op belangrijke punten afgezwakt, en de gewraakte mestboekhouding werd op minder stringente wijze ingevoerd dan Van Aartsen wilde.

De minister slikte de aanpassingen zonder een krimp te geven. Uiterlijk onaangedaan incasseerde hij de kritiek van de Kamer. Dat is typerend voor Van Aartsen. In de vier jaar van zijn ministerschap viel hij niet één keer uit zijn rol van de bedachtzame, afstandelijke bewindsman. Iedereen mag over hem heen vallen, Van Aartsen blijft hoffelijk, een beetje stijf zelfs. Hij herhaalt gewoon zijn standpunten, desnoods tegen beter weten in.

Die houding maakt debatten met de bewindsman saai. Zelfs een erkende zuiger als het Limburgse CDA-kamerlid Van der Linden slaagt er niet in de VVD'er van zijn apropos te brengen. Ook VVD-kamerlid Blauw, die zich al jaren gedraagt alsof hij de enige echte minister van Landbouw is, krijgt steevast vriendelijk antwoord. In het kabinet geldt Van Aartsen als een goede vakminister, zonder veel politieke profilering.
De minister weet wel waarover hij praat, kent zijn dossiers, om een geliefde Haagse terminologie te gebruiken. De herstructurering van de glastuinbouw, het inkrimpen van de varkenshouderij, de overbevissing van de Noordzee of de opwaardering van de Ecologische Hoofdstructuur, Van Aartsen wekt de suggestie dat je hem midden in de nacht wakker kunt maken voor een glashelder referaat.

Hij beheert zijn departement als een goed huisvader, maar komt niet met grootse wetgeving. De omstreden Flora en Faunawet (de opvolger van de jachtwet) die deze week door de Tweede Kamer wordt aangenomen, is een overblijfsel van het vorige kabinet. De mestwet werd grondig uitgekleed door de Kamer. En de begroting voor 1998, toch het oogstjaar van het kabinet, bevatte zo weinig nieuws, dat zijn woordvoerders hem een persconferentie moesten ontraden.

Misschien moet zijn belangrijkste prestatie nog komen. Van Aartsen wil de varkenssector met een kwart inkrimpen. Tot nu toe heeft hij dit idee tegen alle verzet in boven tafel weten te houden. De boerenstand mort, maar Van Aartsengaat stug door met zijn plannen. Sterker nog: hij durft de confrontatie met de boeren aan en bezoekt geregeld agrarische bedrijven.

De herstructurering is het meesterstuk. Als die lukt, wordt zijn ministerschap alsnog een succes. De VVD'er maakt korte metten met jaren CDA-beleid waarin groei van de varkenssector boven milieu en welzijn ging. Van Aartsenverandert dat langzaam. Er is nu meer aandacht voor biologische landbouw en het welzijn van dieren. Dat is een pluspunt. Zonder de sanering van de varkenssector laat de minister evenwel geen grote sporen in het landschap na. Evenmin wacht zijn eventuele opvolger een dikke stapel dossiers met onafgemaakte plannen. Van Aartsen heeft nog acht maanden om zijn taak af te maken.

Annotated correctly: (19) Tweede Kamer; 2x, Van Aartsen; 13x, Europese landbouwbeleid, Noordzee, Ecologische hoofdstructuur, biologische landbouw
Annotated incorrectly: (2) zuiger (Ambiguous term), varkenshouderij (pig farm, linking to pig)
Not annotated: (10) Jozias van Aartsen, de Kamer (Tweede Kamer), gekkekoeienziekte; 2x, varkenspest; 3x, CDA-kamerlid Van der Linden, VVD-kamerlid Blauw, Flora en Faunawet

---

Annotated by the current dutch model

DE MINISTER VAN VIS, vlees, groenten, zee en bos houdt niet van vlees. Hij eet het, maar als hij mag kiezen, neemt hij liever vis. Jozias van Aartsen is daarentegen dol op de natuur en trekt er graag op uit.
Soms neemt hij de fiets om van zijn huis naar het ministerie te gaan. Dat kan goed, want niemand herkent deminister. Zelfs toen de mestwet in de Tweede Kamer werd behandeld en boze boeren het departement belegerden, liet hij een keer de dienstauto staan en kwam hij ongedeerd binnen.

Het ministerschap van Van Aartsen is gedomineerd door de gekkekoeienziekte en de varkenspest. De bewindsman heeft die crises als een behoedzaam manager behandeld.
Hij kondigde bij de BSE direct ingrijpende maatregelen af. Het vlees verdween van de schappen. Bij de varkenspestreageerde de minister trager, voorzichtiger. Het duurde weken voor hij een fokverbod afkondigde. Zijn adviseurswaren woedend over die trage besluitvorming. Van Aartsen verdedigde zich: ingrijpende besluiten moeten zorgvuldig worden genomen. Toch verweet de Kamer hem dat hij achter de feiten aan rende.

De twee veeziekten die het land teisterden, beperkten Van Aartsen in zijn bewegingsvrijheid. Hij had nauwelijks tijdeigen beleid te ontwikkelen. De

gekkekoeienziekte met de BSE-crisis hield hem in 1996 goeddeels van de straat, en dit jaar was het de varkenspest die het gewone regeren onmogelijk maakte.

Het Europese voorzitterschap kostte ook al veel tijd, maar leverde weinig tot de verbeelding sprekende successen op.Dat kan ook haast niet, want het Europese landbouwbeleid is een zaak van lange adem. Snelle succesjes zijnzeldzaam.

Met de mestwet had de minister wel kunnen scoren, maar daar liep het slecht mee af. Van Aartsen trotseerde het boerengeweld buiten de zaal, maar week uiteindelijk voor druk uit de Kamer. De wet werd op belangrijke punten afgezwakt, en de gewraakte mestboekhouding werd op minder stringente wijze ingevoerd dan Van Aartsen wilde.

De minister slikte de aanpassingen zonder een krimp te geven. Uiterlijk onaangedaan incasseerde hij de kritiek van de Kamer. Dat is typerend voor Van Aartsen. In de vier jaar van zijn ministerschap viel hij niet één keer uit zijn rol van debedachtzame, afstandelijke bewindsman. Iedereen mag over hem heen vallen, Van Aartsen blijft hoffelijk, een beetje stijf zelfs. Hij herhaalt gewoon zijn standpunten, desnoods tegen beter weten in.

Die houding maakt debatten met de bewindsman saai. Zelfs een erkende zuiger als het Limburgse CDA-kamerlid Van der Linden slaagt er niet in de VVD'er van zijn apropos te brengen. Ook VVD-kamerlid Blauw, die zich al jaren gedraagt alsof hij de enige echte minister van Landbouw is, krijgt steevast vriendelijk antwoord. In het kabinet geldt Van Aartsen als een goede vakminister, zonder veel politieke profilering.
De minister weet wel waarover hij praat, kent zijn dossiers, om een geliefde Haagse terminologie te gebruiken. Deherstructurering van de glastuinbouw, het inkrimpen van de varkenshouderij, de overbevissing van de Noordzee ofde opwaardering van de Ecologische Hoofdstructuur, Van Aartsen wekt de suggestie dat je hem midden in de nachtwakker kunt maken voor een glashelder referaat.

Hij beheert zijn departement als een goed huisvader, maar komt niet met grootse wetgeving. De omstreden Flora enFaunawet (de opvolger van de jachtwet) die deze week door de Tweede Kamer wordt aangenomen, is een overblijfselvan het vorige kabinet. De mestwet werd grondig uitgekleed door de Kamer. En de begroting voor 1998, toch het oogstjaar van het kabinet, bevatte zo weinig nieuws, dat zijn woordvoerders hem een persconferentie moesten ontraden.

Misschien moet zijn belangrijkste prestatie nog komen. Van Aartsen wil de varkenssector met een kwart inkrimpen. Tot nu toe heeft hij dit idee tegen alle verzet in boven tafel weten te houden. De boerenstand mort, maar VanAartsengaat stug door met zijn plannen. Sterker nog: hij durft de confrontatie met de boeren aan en bezoekt geregeld agrarische bedrijven.

De herstructurering is het meesterstuk. Als die lukt, wordt zijn ministerschap alsnog een succes. De VVD'er maakt korte metten met jaren CDA-beleid waarin groei van de varkenssector boven milieu en welzijn ging. VanAartsenverandert dat langzaam. Er is nu meer aandacht voor biologische landbouw en het welzijn van dieren. Dat is een pluspunt. Zonder de sanering van de varkenssector laat de minister evenwel geen grote sporen in het landschapna. Evenmin wacht zijn eventuele opvolger een dikke stapel dossiers met onafgemaakte plannen. Van Aartsen heeft nog acht maanden om zijn taak af te maken.

Annotated correctly: (14) Jozias van Aartsen, Tweede Kamer; 2x, gekkekoeienziekte; 2x, varkenspest; 3x, BSE, landbouwbeleid, Van der Linden, minister van Landbouw, Noordzee, Ecologische Hoofdstructuur
Annotated incorrectly: (275) DE, vlees, groenten, zee, en, bos, niet, van, vlees, Hij, als, mag, vis, dol, op, de, natuur, en, op, de, fiets, van, zijn, huis, ministerie, Dat, goed, de, minister, de, in, de, en, boeren, departement, keer, de, en, van, de, en, de, De, die, als, manager, Hij, bij, de, BSE, direct, vlees, van, de, Bij, de, de, minister, voor, Zijn, waren, die, de, Kamer, dat, de, feiten, De, die, in, zijn, Hij, tijd, beleid, De, de, crisis, van, de, straat,, en, was, de, die, Europese, voorzitterschap, tijd, de, verbeelding, op, Dat, haast, niet, Europese, zaak, van, lange, zijn, Met, de, de, minister, buiten, de, zaal, week, voor, druk, de, Kamer., De, wet, op, en, de, op, de, Uiterlijk, de, kritiek, van, de, Kamer, Dat, voor, In, de, vier, jaar, van, zijn, niet, keer, zijn, rol, van, de, mag, heen, Hij, zijn, beter, weten, in, Die, de, zuiger, als, niet, in, de, van, zijn, Blauw, die, jaren, de, antwoord, In, kabinet, als, zijn, terminologie, De, van, de, glastuinbouw, van, de, varkenshouderij, de, overbevissing, van, de, de, van, de, de, dat, je, midden, in, de, nacht, voor, Hij, zijn, departement, als, goed, niet, wetgeving, De, Flora, en, de, van, de, die, week, de, van, kabinet., De, de, Kamer., En, de, begroting, voor, van, kabinet, nieuws,, dat, zijn, persconferentie, zijn, wil, de, nu, idee, verzet, in, weten, De, boerenstand, Van, zijn, plannen, de, de, boeren, en, agrarische, bedrijven, De, meesterstuk, Als, die, zijn, De, jaren, beleid, groei, van, de, milieu, en, welzijn, Van, dat, Er, nu, meer, aandacht, voor, biologische, landbouw, en, Dat, de, sanering, van, de, laat, de, minister, in, landschap, zijn, stapel, plannen, maanden, zijn, taak

Not annotated: (10) Van Aartsen; 10x

TEXT 2 - NEWS ARTICLE - Kamer vraagt Schmitz winteronderdak voor Iraniërs in kerkasiel - BINNENLAND - Volkskrant 971217

Annotated by our dutch model

Iraanse asielzoekers die op straat staan of kerkasiel hebben gekregen, moeten van de overheid in de wintermaanden een dak boven hun hoofd krijgen. Dat vindt een meerderheid van de Tweede Kamer. Daarmee komt de Kamer tegemoet aan de kerken die al maanden actievoeren tegen het asielbeleid.

Staatssecretaris Schmitz van Justitie (PvdA) is erop tegen dat uitgeprocedeerde Iraniërs opnieuw opvang krijgen. 'Ze hadden allang moeten teruggaan. Er is ook geen ruimte in de opvang', zei Schmitz vorige week. Donderdag debatteert Schmitz met de Kamer over Iran, maar nu al tekent zich een duidelijke meerderheid af tegen het beleid van Schmitz. Alleen de VVD houdt een slag om de arm.

Kamerlid B. Middel, partijgenoot van Schmitz, vindt dat de staatssecretaris 'een beetje te formalistisch is'. Volgens Middel gaat het om een humaan gebaar. De PvdA'er wil dat 'voor de schrijnende noodgevallen de overheid opvang verzorgt'.

D66-Kamerlid B. Dittrich steunt die gedachte. 'Ik wil in het debat kijken of het voor alle Iraniërs moet gelden, of voor een deel.'

CDA-Kamerlid W. van de Camp had vorige week in een Kamerdebat al aan Schmitz gevraagd om een humaan gebaar te maken. 'Ook de tijd van het jaar speelt een rol', zei Van de Camp.

Er is sinds begin november onduidelijkheid ontstaan over het zogeheten monitorbeleid: het nagaan in Iran of de teruggekeerde asielzoekers wel veilig zijn. Ondanks toezeggingen van Schmitz bleek er niet meer te worden 'gemonitord'.

Vanaf november heeft Schmitz uitzetting van Iraniërs stopgezet in afwachting van duidelijkheid. Schmitz schreef op 5 december dat Iraniërs die voor begin november zijn afgewezen geen opvang krijgen, omdat ze al hadden moeten teruggaan.

De duidelijkheid over de monitoring wordt verschaft in het nieuwe ambtsbericht dat handelt over de veiligheidssituatie in Iran. Dat bericht wordt half januari verwacht. Tot dan is uitzetting niet wenselijk, vinden PvdA,D66, CDA en GroenLinks. Uitzetting laat dan nog tot maart op zich wachten.

De Raad van Kerken vraagt al maanden om aandacht voor de uitgeprocedeerde asielzoekers. Zij moeten Nederland verlaten, maar naar eigen zeggen kunnen ze niet terug. De kerken verlenen hen kerkasiel uit protest tegen het overheidsbeleid.

Vandaag doet de rechtbank in Zwolle uitspraak over een Iraans gezin dat opnieuw in de opvang wil worden genomen. VVD-Kamerlid J. Rijpstra zegt eerst die uitspraak te willen afwachten alvorens een standpunt in te nemen. 'Ik wil ervoor waarschuwen dat we niet te veel uitzonderingen maken. De overheid heeft geen verantwoordelijkheid voor illegalen, uitgeprocedeerde asielzoekers.'

Annotated correctly: (14) asielzoekers; 2x, Tweede Kamer, PvdA; 2x, Iran; 3x, VVD, staatssecretaris, D66, CDA, Groenlinks, Zwolle
Annotated incorrectly: (0)
Not annotated: (29) Iraanse (Iran), de Kamer (Tweede Kamer), Staatssecretaris Schmitz van Justitie, Iraniërs (Iran), Schmitz; 17x, Kamerlid B. Middel, B. Middel, D66-Kamerlid B. Dittrich, B. Dittrich, CDA-Kamerlid W. van de Camp, Van de Camp, De Raad van Kerken, VVD-kamerlid J. Rijpstra

---

Annotated by the current dutch model

Iraanse asielzoekers die op straat staan of kerkasiel hebben gekregen, moeten van de overheid in de wintermaanden een dak boven hun hoofd krijgen. Dat vindt een meerderheid van de Tweede Kamer. Daarmee komt de Kamertegemoet aan de kerken die al maanden actievoeren tegen het asielbeleid.

Staatssecretaris Schmitz van Justitie (PvdA) is erop tegen dat uitgeprocedeerde Iraniërs opnieuw opvang krijgen. 'Ze hadden allang moeten teruggaan. Er is ook geen ruimte in de opvang', zei Schmitz vorige week. Donderdagdebatteert Schmitz met de Kamer over Iran, maar nu al tekent zich een duidelijke meerderheid af tegen het beleid vanSchmitz. Alleen de VVD houdt een slag om de arm.

Kamerlid B. Middel, partijgenoot van Schmitz, vindt dat de staatssecretaris 'een beetje te formalistisch is'. Volgens Middel gaat het om een humaan gebaar. De PvdA'er wil dat 'voor de schrijnende noodgevallen de overheid opvang verzorgt'.

D66-Kamerlid B. Dittrich steunt die gedachte. 'Ik wil in het debat kijken of het voor alle Iraniërs moet gelden, of vooreen deel.'

CDA-Kamerlid W. van de Camp had vorige week in een Kamerdebat al aan Schmitz gevraagd om een humaan gebaarte maken. 'Ook de tijd van het jaar speelt een rol', zei Van de Camp.
Er is sinds begin november onduidelijkheid ontstaan over het zogeheten monitorbeleid: het nagaan in Iran of deteruggekeerde asielzoekers wel veilig zijn. Ondanks toezeggingen van Schmitz bleek er niet meer te worden 'gemonitord'.

Vanaf november heeft Schmitz uitzetting van Iraniërs stopgezet in afwachting van duidelijkheid. Schmitz schreef op 5 december dat Iraniërs die voor begin november zijn afgewezen geen opvang krijgen, omdat ze al hadden moeten teruggaan.

De duidelijkheid over de monitoring wordt verschaft in het nieuwe ambtsbericht dat handelt over deveiligheidssituatie in Iran. Dat bericht wordt half januari verwacht. Tot dan is uitzetting niet wenselijk, vinden PvdA,D66, CDA en GroenLinks. Uitzetting laat dan nog tot maart op zich wachten.

De Raad van Kerken vraagt al maanden om aandacht voor de uitgeprocedeerde asielzoekers. Zij moeten Nederlandverlaten, maar naar eigen zeggen kunnen ze niet terug. De kerken verlenen hen kerkasiel uit protest tegen hetoverheidsbeleid.

Vandaag doet de rechtbank in Zwolle uitspraak over een Iraans gezin dat opnieuw in de opvang wil worden genomen. VVD-Kamerlid J. Rijpstra zegt eerst die uitspraak te willen afwachten alvorens een standpunt in te nemen. 'Ik wil ervoor waarschuwen dat we niet te veel uitzonderingen maken. De overheid heeft geen verantwoordelijkheidvoor illegalen, uitgeprocedeerde asielzoekers.'

Annotated correctly: (17) Iraanse, asielzoekers; 3x, Tweede Kamer, Iran; 3x, Staatssecretaris; 2x, Iraniërs; 2x, VVD, D66, GroenLinks, Raad van Kerken, Zwolle
Annotated incorrectly: (154) die, op, straat, kerkasiel, van, de, overheid, in, de, dak, hun, hoofd, Dat, meerderheid, van, de, de, Kamer, de, die, maanden, van, Justitie, dat, Er, ruimte, in, de, week., Donderdag, de, Kamer, nu, meerderheid, beleid, van, de, slag, de, arm, Kamerlid, B, van, dat, de, gebaar., De, wil, dat, voor, de, de, overheid, D66, Kamerlid, B, Dittrich, die, gedachte, Ik, wil, in, debat, voor, Iraniërs, voor, deel, Kamerlid, van, de, Camp, week, in, Kamerdebat, gebaar, de, tijd, van, jaar, rol, Er, begin, november, in, de, veilig, zijn, van, bleek, niet, meer, november, van, in, van, schreef, op, 5, december, dat, die, voor, begin, november, zijn, De, de, in, ambtsbericht, dat, de, in, Dat, bericht, januari, niet, en, laat, maart, op, De, maanden, aandacht, voor, de, Nederland, niet, De, kerkasiel, protest, overheidsbeleid, Vandaag, de, rechtbank, in, Iraans, gezin, dat, in, de, wil, VVD-Kamerlid, J, Rijpstra, die, standpunt, in, Ik, wil, dat, niet, De, overheid, voor
Not annotated: (11) Schmitz; 7x, PvdA; 2x, Van de Camp, CDA

TEXT 3 - NEWS ARTICLE - Clinton moet Amerikanen uit hun auto lokken - BUITENLAND - Volkskrant 970630

Annotated by our dutch model

De teleurstelling over 'het verraad van de beloftes van Rio' was aan het einde van de 'Aardetop plus 5' groot, althans onder een deel van de milieubeweging en de ontwikkelingslanden. De week lange speciale vergadering van deVerenigde Naties in New York leverde niet eens een politieke slotverklaring op.

De verdeeldheid tussen arm en rijk, noord en zuid, oost en west over de uitvoering van het actieprogramma voor een schoner milieu en duurzame groei, was te groot om in één week te overbruggen. De meningsverschillen over klimaatverandering, financiering, water en bossen liepen dwars door de continenten, de geïndustrialiseerde wereld en de ontwikkelingslanden.
Er is niet een bepaald land of een groep van landen die verantwoordelijk kan worden gesteld voor de mislukking van de VN-bijeenkomst . Het oplossen van wereldwijde milieu- en ontwikkelingsproblemen is daarvoor te gecompliceerd en bevat te veel politieke en economische dilemma's.

De sombere commentaren waarmee het milieu-establishment (diplomaten, ministers, actiegroepen, sommige milieujournalisten) New York verliet was verklaarbaar, maar hield te weinig rekening met de opmerkelijke toespraak van president Clinton. Voor het eerst sinds de grote VN-milieu-top van 1992 in Rio de Janeiro heeft een Amerikaanse president zich uitgesproken voor bindende limieten aan de uitstoot van broeikasgassen.

Deze gassen, vooral kooldioxide en methaan, veroorzaken opwarming van het klimaat , hetgeen grote gevolgen heeft voor ecosystemen, landbouw- en klimaatzones en de zeeën.
Onder druk van de Europese Unie, van zijn eigen vice-president Al Gore en van de invloedrijke Amerikaanse milieubeweging heeft Clinton toegezegd dat er nog dit jaar 'bindende en realistische limieten' afgesproken moeten worden. Dat moet gebeuren in het Japanse Kyoto, waar de landen die de Conventie over Klimaatverandering hebben ondertekend in december bijeenkomen.
Aangezien de VS met een uitstoot van ruim 5000 ton kooldioxide per jaar een van de grootste vervuilers is, zal het Clinton de grootste moeite kosten zijn toezegging in amper zes maanden waar te maken. Hij heeft zich vastgelegd op een indrukwekkende doelstelling - overigens zonder percentages en data te noemen.

Minstens zo belangrijk was deze week dat vice-president Gore op de openingsdag van deze Aardetop al een voorzet gaf. Als mogelijke opvolger van Clinton zal het wellicht Gore zijn die de afspraken van Kyoto moet uitvoeren, ondanks het verzet van het Amerikaanse bedrijfsleven. De grote nutsbedrijven, de olieproducenten, de autoindustrie en de chemische industrie vormen ieder voor zich, maar zeker gezamenlijk, een lobby die het Congres niet makkelijk kan negeren.

In 1993 torpedeerde een coalitie van Republikeinen en conservatieve Democraten een bescheiden plan van Clinton en Gore voor een 'energiebelasting' om de uitstoot van kooldioxide te verminderen. Dat gevecht zal zich op een aanzienlijke grotere schaal herhalen als Clinton in Kyoto afspraken maakt die werkelijk gevolgen hebben voor de wijze waarop in Amerika energie wordt verbruikt.
Daarom begint Clinton op korte termijn met een campagne om de Amerikaanse bevolking de aard van het tamelijk abstracte probleem uit leggen. Zonder een dergelijke campagne, waarin hij de kosten en opofferingen niet zal kunnen negeren, is het ondenkbaar dat het Congres de Klimaatconventie van Kyoto zal steunen.

In een land dat verslaafd is aan goedkope fossiele brandstoffen en comfortabele auto's is dat geen kleine opgave. Clinton moet bovendien het land, het Congres en het bedrijfsleven ervan overtuigen dat er meer moet worden geïnvesteerd in nieuwe, schonere technologie. Milieukosten, moet hij uitleggen, zijn een onderdeel van de productiekosten en geen vorm van verkapte belastingen.
Op internationaal vlak zijn de uitdagingen niet minder groot. De opwarming van de aarde kan alleen worden voorkomen als India, China, Japan, de Oost-Aziatische tijgers, de olieproducerende landen, maar ook Turkije, Spanje en Italië volledig meewerken.

In New York is gebleken hoe terughoudend deze landen zijn. Geen enkel land wil zijn economische groei op het spel zetten voor de bestrijding van een probleem dat zich pas in de tweede helft van de volgende eeuw zal manifesteren.

Al deze thema's heeft Clinton in overleg met zijn mogelijke opvolger Al Gore op de agenda van zijn tweede termijn geplaatst. Grote woorden in New York dienen gevolgd te worden door grote daden in Kyoto. Pas dan kan worden beoordeeld of de beloftes van Rio zijn verraden.

Annotated correctly: (18) ontwikkelingslanden; 2x, Verenigde Naties, klimaatverandering, broeikasgassen, methaan, ecosystemen, Europese Unie, Al Gore; 2x, Republikeinen, kooldioxide, India, China, Japan, Turkije, Spanje, Italië
Annotated incorrectly: (3) milieubeweging; 2x, fossiele brandstoffen
Not annotated: (24) New York; 4x, Clinton; 9x, Rio de Janeiro, Kyoto; 5x, VS, Gore; 3x (Al Gore), Democraten

---

Annotated by the current dutch model

De teleurstelling over 'het verraad van de beloftes van Rio' was aan het einde van de 'Aardetop plus 5' groot, althans onder een deel van de milieubeweging en de ontwikkelingslanden. De week lange speciale vergadering vandeVerenigde Naties in New York leverde niet eens een politieke slotverklaring op.

De verdeeldheid tussen arm en rijk, noord en zuid, oost en west over de uitvoering van het actieprogramma voor een schoner milieu en duurzame groei, was te groot om in één week te overbruggen. De meningsverschillen overklimaatverandering, financiering, water en bossen liepen dwars door de continenten, de geïndustrialiseerde wereld en de ontwikkelingslanden.
Er is niet een bepaald land of een groep van landen die verantwoordelijk kan worden gesteld voor de mislukking van de VN-bijeenkomst . Het oplossen van wereldwijde milieu- en ontwikkelingsproblemen is daarvoor te gecompliceerd en bevat te veel politieke en economische dilemma's.

De sombere commentaren waarmee het milieu-establishment (diplomaten, ministers, actiegroepen, sommige milieujournalisten) New York verliet was verklaarbaar, maar hield te weinig rekening met de opmerkelijke toespraakvan president Clinton. Voor het eerst sinds de grote VN-milieu-top van 1992 in Rio de Janeiro heeft een Amerikaanse president zich uitgesproken voor bindende limieten aan de uitstoot van broeikasgassen.

Deze gassen, vooral kooldioxide en methaan, veroorzaken opwarming van het klimaat , hetgeen grote gevolgen heeft voor ecosystemen, landbouw- en klimaatzones en de zeeën.
Onder druk van de Europese Unie, van zijn eigen vice-president Al Gore en van de invloedrijke Amerikaansemilieubeweging heeft Clinton toegezegd dat er nog dit jaar 'bindende en realistische limieten' afgesproken moeten worden. Dat moet gebeuren in het Japanse Kyoto, waar de landen die de Conventie over Klimaatverandering hebben ondertekend in december bijeenkomen.
Aangezien de VS met een uitstoot van ruim 5000 ton kooldioxide per jaar een van de grootste vervuilers is, zal het Clinton de grootste moeite

kosten zijn toezegging in amper zes maanden waar te maken. Hij heeft zich vastgelegd opeen indrukwekkende doelstelling - overigens zonder percentages en data te noemen.

Minstens zo belangrijk was deze week dat vice-president Gore op de openingsdag van deze Aardetop al een voorzet gaf. Als mogelijke opvolger van Clinton zal het wellicht Gore zijn die de afspraken van Kyoto moet uitvoeren, ondanks het verzet van het Amerikaanse bedrijfsleven. De grote nutsbedrijven, de olieproducenten, de autoindustrie en de chemische industrie vormen ieder voor zich, maar zeker gezamenlijk, een lobby die het Congres niet makkelijk kan negeren.

In 1993 torpedeerde een coalitie van Republikeinen en conservatieve Democraten een bescheiden plan van Clinton en Gore voor een 'energiebelasting' om de uitstoot van kooldioxide te verminderen. Dat gevecht zal zich op een aanzienlijke grotere schaal herhalen als Clinton in Kyoto afspraken maakt die werkelijk gevolgen hebben voor de wijze waarop in Amerika energie wordt verbruikt.
Daarom begint Clinton op korte termijn met een campagne om de Amerikaanse bevolking de aard van het tamelijk abstracte probleem uit leggen. Zonder een dergelijke campagne, waarin hij de kosten en opofferingen niet zal kunnen negeren, is het ondenkbaar dat het Congres de Klimaatconventie van Kyoto zal steunen.

In een land dat verslaafd is aan goedkope fossiele brandstoffen en comfortabele auto's is dat geen kleine opgave.Clinton moet bovendien het land, het Congres en het bedrijfsleven ervan overtuigen dat er meer moet worden geïnvesteerd in nieuwe, schonere technologie. Milieukosten, moet hij uitleggen, zijn een onderdeel van de productiekosten en geen vorm van verkapte belastingen.
Op internationaal vlak zijn de uitdagingen niet minder groot. De opwarming van de aarde kan alleen worden voorkomen als India, China, Japan, de Oost-Aziatische tijgers, de olieproducerende landen, maar ook Turkije, Spanje en Italië volledig meewerken.

In New York is gebleken hoe terughoudend deze landen zijn. Geen enkel land wil zijn economische groei op het spelzetten voor de bestrijding van een probleem dat zich pas in de tweede helft van de volgende eeuw zal manifesteren.

Al deze thema's heeft Clinton in overleg met zijn mogelijke opvolger Al Gore op de agenda van zijn tweede termijn geplaatst. Grote woorden in New York dienen gevolgd te worden door grote daden in Kyoto. Pas dan kan worden beoordeeld of de beloftes van Rio zijn verraden.

Annotated correctly: (19) ontwikkelingslanden; 2x, New York; 4x, VN; 2x, Al Gore, Clinton, Republikeinen, Amerikaanse, Fossiele Brandstoffen, Gore, India, China, Japan, Turkije, Spanje
Annotated incorrectly: (200) de; 43x, van; 33x, was; 4x, milieubeweging, week; 3x, vergadering, niet; 5x, op; 7x, arm, rijk, noord, zuid, oost, west, voor; 7x, milieu; 4x, groei, in; 14x, klimaatverandering, financiering, water, dwars, land; 3x, economische, groep, dilemma's, establishment, actiegroepen, rekening, toespraak, top, Amerikaanse President, uitstoot, broeikasgassen, kooldioxide; 2x, jaar, maanden, waar; 2x, vice-president, verzet, bedrijfsleven; 2x, autoindustrie, chemische industrie, lobby, coalitie, Democraten, conservatieve, energiebelasting, uitstoot, gevecht, schaal, energie, campagne; 2x, bevolking, verslaafd, technologie, productiekosten, vorm, internationaal, groot, opwarming, aarde, tijgers, als; 3x, enkel, wil, zijn; 7x, spel, eeuw, overleg, woorden, pas; 2x, Rio
Not annotated: (12) Verenigde Naties, Rio de Janero, Italie, VS, Kyoto; 5x, Amerika, Amerikaanse, Oost-Aziatische
---

TEXT 4 - MOVIE REVIEW - Stolen - Simon West - Nu Filmrecensies 130116

Annotated by our dutch model

In de wisselvallige acteercarrière van Nicolas Cage kan de uitdrukking 'Going back to wrong' heel veel betekenen. In Stolen pakt 'wrong' nog best leuk uit.

New Orleans. Vier uur 's ochtends. Voor de diamantbeurs staat het busje van inbrekers Will Montgomery (Nicolas Cage), Vincent (Josh Lucas), Hoyt (M.C. Gainey) en Riley (Malin Åkerman). Ze hebben niet door dat de politie ze in de gaten houdt. Toch is inspecteur Harlend (Danny Huston) er niet gerust op: "Too easy. This crew doesn't do easy."

Tien miljoen
Will Montgomery is namelijk een ontzettend gewiekste dief, al loopt deze inbraak helaas slecht voor hem af. Nadat ze 10 miljoen dollar hebben buitgemaakt, krijgen Will en Vincent slaande ruzie. Resultaat: de zwaargewonde Vincent en de anderen laten Will en het geld achter terwijl de politiesirenes naderen.

Will draait dus in zijn eentje de bak in. Wanneer hij 8 jaar later vrijkomt zijn z'n oude kompanen in geen velden of wegen te bekennen, is hij vervreemd van zijn dochter Alison (Sami Gayle) en is de politie nog steeds niet klaar met hem. Want, waar is destijds die 10 miljoen dollar gebleven?

Niet alleen inspecteur Harlend weet zeker dat Will dat geld ergens verstopt heeft; ook zijn oude maat Vincent denkt dat. De engerd heeft tegenwoordig een kunstbeen en klust bij als taxichauffeur. Hij ontvoert Alison en geeft Will precies 12 uur de tijd om met het geld over de brug te komen. Will moet, kortom, 'back to wrong'.

Guilty pleasure
Stolen is gemaakt door Simon West en dat geeft al een aardige indicatie van wat je te wachten staat. De regisseur debuteerde in 1997 met Con Air. Verleden jaar verzamelde hij de bekendste krachtpatsers van de internationale actiefilm voor The Expendables 2, nog zo'n 'guilty pleasure'. Origineel noch verrassend dus, maar Stolen komt wel over de brug met een aantal vlotte actiescènes. Erg waarschijnlijk is het allemaal ook niet. Zou je in Amerika werkelijk eerder vrijkomen als ze de opbrengst van je bankroof niet kunnen terugvinden? En hoe snel smelt goud eigenlijk?

Yellow cab
Cage, die toch op zijn best is als hij krankzinnige types mag spelen, zet zich niet heel erg in voor deze film of de lachwekkend dramatische bijna-finale. Tegenspeler Josh Lucas geeft echter wél vol gas. Vieze Vincent is het soort taxichauffeur waar geen enkel weldenkend mens bij zou instappen.
Die arme Alison ligt bijna de hele film gekneveld in de achterbak van Vincents 'yellow cab'. En probeer die dan maar eens terug te vinden tussen de duizenden gele taxi's in New York! Een perfecte verstopplek, midden in het zicht. Zo'n ideetje maakt Stolen - eigenlijk net als de rest van Wests oeuvre - een pretfilm tegen beter weten in.

Annotated correctly: (8) Nicolas Cage; 2x, New Orleans, Josh Lucas; 2x, Danny Huston, Simon West, Con Air
Annotated incorrectly: (1) The Expendables (The Expendables 2)
Not annotated: (9) Stolen; 4x, M.C. Gainey, Malin Akerman, Sami Gayle, Amerika, Cage (Nicolas Cage)

Annotated by the current dutch model

In de wisselvallige acteercarrière van Nicolas Cage kan de uitdrukking 'Going back to wrong' heel veel betekenen. In Stolen pakt 'wrong' nog best leuk uit.

New Orleans. Vier uur 's ochtends. Voor de diamantbeurs staat het busje van inbrekers Will Montgomery (Nicolas Cage), Vincent (Josh Lucas), Hoyt (M.C. Gainey) en Riley (Malin Åkerman). Ze hebben niet door dat de politie ze in degaten houdt. Toch is inspecteur Harlend (Danny Huston) er niet gerust op: "Too easy. This crew doesn't do easy."

Tien miljoen
Will Montgomery is namelijk een ontzettend gewiekste dief, al loopt deze inbraak helaas slecht voor hem af. Nadat ze 10 miljoen dollar hebben buitgemaakt, krijgen Will en Vincent slaande ruzie. Resultaat: de zwaargewonde Vincent endе anderen laten Will en het geld achter terwijl de politiesirenes naderen.

Will draait dus in zijn eentje de bak in. Wanneer hij 8 jaar later vrijkomt zijn z'n oude kompanen in geen velden of wegen te bekennen, is hij vervreemd van zijn dochter Alison (Sami Gayle) en is de politie nog steeds niet klaar met hem. Want, waar is destijds die 10 miljoen dollar gebleven?

Niet alleen inspecteur Harlend weet zeker dat Will dat geld ergens verstopt heeft; ook zijn oude maat Vincent denkt dat. De engerd heeft tegenwoordig een kunstbeen en klust bij als taxichauffeur. Hij ontvoert Alison en geeft Will precies 12 uur de tijd om met het geld over de brug te komen. Will moet, kortom, 'back to wrong'.

Guilty pleasure
Stolen is gemaakt door Simon West en dat geeft al een aardige indicatie van wat je te wachten staat. De regisseur debuteerde in 1997 met Con Air. Verleden jaar verzamelde hij de bekendste krachtpatsers van de internationale actiefilm voor The Expendables 2, nog zo'n 'guilty pleasure'. Origineel noch verrassend dus, maar Stolen komt wel over de brug met een aantal vlotte actiescènes. Erg waarschijnlijk is het allemaal ook niet. Zou je in Amerika werkelijk eerder vrijkomen als ze de opbrengst van je bankroof niet kunnen terugvinden? En hoe snel smelt goud eigenlijk?

Yellow cab
Cage, die toch op zijn best is als hij krankzinnige types mag spelen, zet zich niet heel erg in voor deze film of de lachwekkend dramatische bijna-finale. Tegenspeler Josh Lucas geeft echter wél vol gas. Vieze Vincent is het soort taxichauffeur waar geen enkel weldenkend mens bij zou instappen.
Die arme Alison ligt bijna de hele film gekneveld in de achterbak van Vincents 'yellow cab'. En probeer die dan maar eens terug te vinden tussen de duizenden gele taxi's in New York! Een perfecte verstopplek, midden in het zicht. Zo'n ideetje maakt Stolen - eigenlijk net als de rest van Wests oeuvre - een pretfilm tegen beter weten in.

Annotated correctly: (1) New York
Annotated incorrectly: (57) de; 45x, 's; 2x, politie; 3x, 't, dollar; 2x, geld; 3x, goud
Not annotated: (18) Nicolas Cage; 2x, Stolen; 4x, New Orleans, John Lucas; 2x, M.C. Gainey, Malin Akerman, Danny Huston, Sami Gayle, Simon West, Con Air, The Expendables 2, Amerika, Cage (Nicolas Cage)

TEXT 5 - MOVIE REVIEW - Lawless - John Hillcoat - Nu Filmrecensies 121114

Annotated by our dutch model

Misdaadfilm spelend in de tijd van de Amerikaanse drooglegging in de jaren dertig. Met Tom Hardy en Shia LaBeouf als illegale drankstokers

De periode van de drooglegging biedt al tachtig jaar stof voor films en tv-series.

Al in de jaren dertig van de vorige eeuw leverde dat de gangsterfilms van Warner Bros op, zoals Public Enemy, en later kwamen daar tv-series als The Intouchables bij, dat ook werd verfilmd. Een serie als Boardwalk Empire toont aan dat de creatieve bron van de drooglegging nog niet opgedroogd is.

De meeste gangsterfilms en -series spelen zich echter in de grote stad af, waar de illegale drank werd afgezet, en zelden op de plek waar het werd gemaakt. Lawless richt zich juist wel op het platteland waar de 'moonshine' werd gestookt, en op drie broers die met bootlegging hun geld verdienen: de Bondurant broers, die ook echt hebben bestaan.

Drie broers
De zwijgzame Forrest Bondurant (Tom Hardy) is de leider van het trio, de oorlogsveteraan Howard (Jason Clarke) de dommekracht en de naïeve Jack is de jongste broer die zo graag voor vol aangezien wil worden. Met hun illegale zaakjes gaat het goed, zolang ze zich maar gedeisd houden en de juiste agenten smeergeld betalen.
Maar de komst van special agent Charlie Rakes (Guy Pearce) dreigt een streep door de rekening te zetten. De bizar uitgedoste wetsdienaar doet zich voor als onkreukbaar en weigert smeergeld van de bootleggers aan te nemen. Hij wil persoonlijk afrekenen met de Bondurant broers, ook als hij daarvoor zelf de wet moet overtreden.

Vrouwelijk schoon
Ondertussen gaat het dagelijks leven door: de fraaie Maggie (Jessica Chastain) dringt zich subtiel op aan Forrest terwijl Jack werk probeert te maken van de puriteinse Bertha (Mia Wasikowska). Maar laten de bootleggers zich niet te veel afleiden door dit vrouwelijk schoon? En wat voor listen heeft Rakes nog in petto?

Lawless, naar een script van zanger Nick Cave, is een harde maar sfeervolle gangsterfilm. Het geflirt van Shia Labeouf hadden we kunnen missen, maar Tom Hardy maakt als zijn zwijgzame broer veel goed. En de soundtrack - metcountryversies van White Light/White Heat van The Velvet Underground - draagt nog verder bij aan de sfeer.

Annotated correctly: (16) Amerikaanse drooglegging, Tom Hardy; 3x, Shia LaBeouf, Public Enemy, Boardwalk Empire, Jason Clarke, guy Pearce, bootleggers; 2x, Mia Wasikowska, Nick Cace, countryversies, White Light/White Heat, The Velvet Underground
Annotated incorrectly: (0)
Not annotated: (7) drooglegging (Amerikaanse drooglegging), Warner Bros, The Intouchables, Lawless, moonshine, bootlegging, Jessica Chastain, Shia LaBeouf, oorlogsveteraan, soundtrack

---

Annotated by the current dutch model

Misdaadfilm spelend in de tijd van de Amerikaanse drooglegging in de jaren dertig. Met Tom Hardy en Shia LaBeouf als illegale drankstokers

De periode van de drooglegging biedt al tachtig jaar stof voor films en tv-series.

Al in de jaren dertig van de vorige eeuw leverde dat de gangsterfilms van Warner Bros op, zoals Public Enemy, en later kwamen daar tv-series als The Intouchables bij, dat ook werd verfilmd. Een serie als Boardwalk Empire toont aan datde creatieve bron van de drooglegging nog niet opgedroogd is.

De meeste gangsterfilms en -series spelen zich echter in de grote stad af, waar de illegale drank werd afgezet, en zelden op de plek waar het werd gemaakt. Lawless richt zich juist wel op het platteland waar de 'moonshine' werd gestookt, en op drie broers die met bootlegging hun geld verdienen: de Bondurant broers, die ook echt hebben bestaan.

Drie broers
De zwijgzame Forrest Bondurant (Tom Hardy) is de leider van het trio, de oorlogsveteraan Howard (Jason Clarke) dedommekracht en de naïeve Jack is de jongste broer die zo graag voor vol aangezien wil worden. Met hun illegale zaakjes gaat het goed, zolang ze zich maar gedeisd houden en de juiste agenten smeergeld betalen.
Maar de komst van special agent Charlie Rakes (Guy Pearce) dreigt een streep door de rekening te zetten. De bizar uitgedoste wetsdienaar doet zich voor als onkreukbaar en weigert smeergeld van de bootleggers aan te nemen. Hijwil persoonlijk afrekenen met de Bondurant broers, ook als hij daarvoor zelf de wet moet overtreden.

Vrouwelijk schoon
Ondertussen gaat het dagelijks leven door: de fraaie Maggie (Jessica Chastain) dringt zich subtiel op aan Forrest terwijl Jack werk probeert te maken van de puriteinse Bertha (Mia Wasikowska). Maar laten de bootleggers zich niet te veel afleiden door dit vrouwelijk schoon? En wat voor listen heeft Rakes nog in petto?

Lawless, naar een script van zanger Nick Cave, is een harde maar sfeervolle gangsterfilm. Het geflirt van Shia Labeouf hadden we kunnen missen, maar Tom Hardy maakt als zijn zwijgzame broer veel goed. En de soundtrack - metcountryversies van White Light/White Heat van The Velvet Underground - draagt nog verder bij aan de sfeer.

Annotated correctly: (2) oorlogsveteraan, soundtrack
Annotated incorrectly: (52) in; 4x, de; 29x, tijd, Met; 2x, Al, verfilmd, bij, stad, waar; 3x, op; 4x, agent, Charlie, wil; 2x, wet, hun; 2x, Jack, missen
Not annotated: (24) Amerikaanse drooglegging, Tom Hardy; 3x, Shia LaBeouf, Public Enemy, Boardwalk Empire, Jason Clarke, guy Pearce, bootleggers; 2x, Mia Wasikowska, Nick Cace, countryversies, White Light/White Heat, The Velvet Underground, drooglegging (Amerikaanse drooglegging), Warner Bros, The Intouchables, Lawless, moonshine, bootlegging, Jessica Chastain, Shia LaBeouf

---

TEXT 6 - MOVIE REVIEW - Looper - Rian Johnson - Nu Filmrecensies 121128

Annotated by our dutch model

Ingenieuze tijdreisfilm waarin Joseph Gordon-Levitt en Bruce Willis als oude en jonge versie van dezelfde persoon elkaar tegen het lijf lopen.

Het is op deze plek al vaker betoogd: tijdreizen in films blijft tricky. Tegen elke geslaagde tijdreisfilm (Back to the Future, Terminator) staat een minder geslaagde (Back to the Future 2, Terminator 3).

De sf-film Looper pakt weer geslaagd uit, omdat het tijdreizen slechts een van de elementen is in een goed verhaal met gelaagde personages en fraai uitgevoerde actiescènes.

2044/2074
Looper speelt in 2044, een tijd waarin de kloof tussen arm en rijk verder is vergroot. Joe (Joseph Gordon-Levitt, uitInception en The Dark Knight Rises) behoort tot de rijke klasse: hij is een goedbetaalde 'looper', een huurmoordenaar die criminelen executeert die hij weer uit zijn toekomst (namelijk 2074) krijgt opgestuurd.
De tijdssprong die de slachtoffers maken heet een loop, vandaar de titel.

Jonge Joe vs Oude Joe
Joe leidt een luxe maar leeg leventje, vol retro-verzetjes (pick up!) en drugs die als oogdruppels dienen te worden toegediend. Hij krijgt een wake-up call wanneer zijn volgende slachtoffer uit de toekomst een goede bekende van hem is, namelijk een 30 jaar oudere versie van hemzelf (Bruce Willis). Jonge Joe aarzelt iets te lang, en oude Joe ontsnapt.

Daarmee laadt jonge Joe zich de woede van zijn opdrachtgever op zich. Want wie weigert de oudere versie van zichzelf uit de weg te ruimen wacht een vreselijk lot, zagen we eerder al bij een vriend en collega van Joe (rol van Paul Dano). Nu moet Joe zijn luxe leventje verlaten, onderduiken én

proberen zijn oude versie alsnog op te sporen.

Dilemma
Looper begeeft zich op glad ijs met het tijdreisprincipe: elke wijziging in het nu heeft gevolgen voor de toekomst.
Maar regisseur Rian Johnson, die ook het slimme script schreef, komt er goed mee weg. Met als voornaamste troef Gordon-Levitt (die ook in Johnsons spelfilmdebuut Brick de hoofdrol speelde).

Want met zijn scheve glimlach en fluisterstem, gecompleteerd door een neusprothese, lijkt Gordon-Levitt bewonderenswaardig veel op een jonge Bruce Willis.

Dat persoonlijke element geeft het kat-en-muis-spel in het verhaal een diepere laag, en ook de rest van de film zit vol slimme vondsten en sterke actiescènes. Looper is een zeer geslaagde sf-film waarbij het publiek ook nog eens de hersens mag gebruiken.

Annotated correctly: (9) Joseph Gordon-Levitt; 2x, Bruce Willis; 3x, tijdreizen, Inception, The Dark Knight Rises, Rian Johnson
Annotated incorrectly: (0)
Not annotated: (11) tijdreizen, Back to the Future, Terminator, Back to the Future 2, Terminator 3, Looper; 4x, Paul Dano, Gordon-Levitt, Brick

---

Annotated by the current dutch model

Ingenieuze tijdreisfilm waarin Joseph Gordon-Levitt en Bruce Willis als oude en jonge versie van dezelfde persoon elkaar tegen het lijf lopen.

Het is op deze plek al vaker betoogd: tijdreizen in films blijft tricky. Tegen elke geslaagde tijdreisfilm (Back to the Future, Terminator) staat een minder geslaagde (Back to the Future 2, Terminator 3).

De sf-film Looper pakt weer geslaagd uit, omdat het tijdreizen slechts een van de elementen is in een goed verhaal met gelaagde personages en fraai uitgevoerde actiescènes.

2044/2074
Looper speelt in 2044, een tijd waarin de kloof tussen arm en rijk verder is vergroot. Joe (Joseph Gordon-Levitt, uit Inception en The Dark Knight Rises) behoort tot de rijke klasse: hij is een goedbetaalde 'looper', een huurmoordenaar die criminelen executeert die hij weer uit zijn toekomst (namelijk 2074) krijgt opgestuurd.
De tijdssprong die de slachtoffers maken heet een loop, vandaar de titel.

Jonge Joe vs Oude Joe
Joe leidt een luxe maar leeg leventje, vol retro-verzetjes (pick up!) en drugs die als oogdruppels dienen te worden toegediend. Hij krijgt een wake-up call wanneer zijn volgende slachtoffer uit de toekomst een goede bekende van hem is, namelijk een 30 jaar oudere versie van hemzelf (Bruce Willis). Jonge Joe aarzelt iets te lang, en oude Joe ontsnapt.

Daarmee laadt jonge Joe zich de woede van zijn opdrachtgever op zich. Want wie weigert de oudere versie van zichzelf uit de weg te ruimen wacht een vreselijk lot, zagen we eerder al bij een vriend en collega van Joe (rol van Paul Dano). Nu moet Joe zijn luxe leventje verlaten, onderduiken én proberen zijn oude versie alsnog op te sporen.

Dilemma
Looper begeeft zich op glad ijs met het tijdreisprincipe: elke wijziging in het nu heeft gevolgen voor de toekomst.
Maar regisseur Rian Johnson, die ook het slimme script schreef, komt er goed mee weg. Met als voornaamste troefGordon-Levitt (die ook in Johnsons spelfilmdebuut Brick de hoofdrol speelde).

Want met zijn scheve glimlach en fluisterstem, gecompleteerd door een neusprothese, lijkt Gordon-Levitt bewonderenswaardig veel op een jonge Bruce Willis.

Dat persoonlijke element geeft het kat-en-muis-spel in het verhaal een diepere laag, en ook de rest van de film zit vol slimme vondsten en sterke actiescènes. Looper is een zeer geslaagde sf-film waarbij het publiek ook nog eens dehersens mag gebruiken.

Annotated correctly: (4) tijdreizen; 2x, sf; 2x
Annotated incorrectly: (15) Gordon, de; 13x, vs
Not annotated: (19) Joseph Gordon-Levitt; 2x, Bruce Willis; 3x, Inception, The Dark Knight Rises, Rian Johnson, Back to the Future, Terminator, Back to the Future 2, Terminator 3, Looper; 4x, Paul Dano, Gordon-Levitt, Brick

---

TEXT 7 - TOURIST INFORMATION - Tips: Beste Pubs in Londen - reistipseuropa.nl 110531

Annotated by our dutch model

Het is onmogelijk om London los te zien van de pubs. Bij een bezoek aan London valt het pas op hoeveel mooie pubs de stad rijk is, die (niet geheel onbelangrijk) veel lokaal bier aanbieden. In dit artikel vier van de beste pubs in London.

The Mayflower
The Mayflower staat bekend om haar historie. Een pub heeft al op dezelfde plek gestaan sinds 1620 ten tijden van de Pilgrims. De huidige pub is gebouwd in de 18e eeuw en is een typische Engelse pub met grote houten balken enhouten vloeren. Op de eerste verdieping vindt je een (duur) restaurant met uitzicht op de Themes.

Zeitgeist London
Een hele gok om een Duitse pub in London te beginnen. Niet alleen heeft de pub een Duitse naam, ze hebben Duitsbier op de tap, Duitse medewerkers en Duits voetbal wordt hier op TV's vertoond. Toch trekt Zeitgeist niet alleen veel buitenlandse mensen. Ook de lokale bevolking

weet de pub te waarderen.

Sir Richard Steel
Vernoemd naar de mede-oprichter van The Spectator. Het is een nogal excentrieke pub, met verkeersborden enopgezette dieren aan de muur. In deze pub wordt vooral ale gedronken, met vier verschillende soorten bier op detap. Naast bier kun je hier ook goedkoop Thais eten, en er zijn geregeld optredens.

Greenwich Union
Als je de wat meer populaire pubs mijdt en je zoekt wat verder kom je bij de Greenwich Union. Het is een kleine, smalle pub die vooral populair is bij de jeugd. Naast veel soorten bier zijn er ook wat minder bekende bieren te vinden, zoals chocolate- en raspberrybeer. Mocht je een vervente Internetter zijn, je kunt hier gratis gebruik maken van Wi-Fi.

Annotated correctly: (1) Thais
Annotated incorrectly: (5) bier; 3x (not important enough to annotate), Mayflower (wrong mayflower), The Spectator
Not annotated: (13) London; 4x (due to spelling mistake), Pilgrims, 18e eeuw, Engelse, Themes, Duitse; 5x

---

Annotated by the current dutch model

Het is onmogelijk om London los te zien van de pubs. Bij een bezoek aan London valt het pas op hoeveel mooie pubs de stad rijk is, die (niet geheel onbelangrijk) veel lokaal bier aanbieden. In dit artikel vier van de beste pubs in London.

The Mayflower
The Mayflower staat bekend om haar historie. Een pub heeft al op dezelfde plek gestaan sinds 1620 ten tijden van de Pilgrims. De huidige pub is gebouwd in de 18e eeuw en is een typische Engelse pub met grote houten balken enhouten vloeren. Op de eerste verdieping vindt je een (duur) restaurant met uitzicht op de Themes.

Zeitgeist London
Een hele gok om een Duitse pub in London te beginnen. Niet alleen heeft de pub een Duitse naam, ze hebben Duitsbier op de tap, Duitse medewerkers en Duits voetbal wordt hier op TV's vertoond. Toch trekt Zeitgeist niet alleen veel buitenlandse mensen. Ook de lokale bevolking weet de pub te waarderen.

Sir Richard Steel
Vernoemd naar de mede-oprichter van The Spectator. Het is een nogal excentrieke pub, met verkeersborden enopgezette dieren aan de muur. In deze pub wordt vooral ale gedronken, met vier verschillende soorten bier op detap. Naast bier kun je hier ook goedkoop Thais eten, en er zijn geregeld optredens.

Greenwich Union
Als je de wat meer populaire pubs mijdt en je zoekt wat verder kom je bij de Greenwich Union. Het is een kleine, smalle pub die vooral populair is bij de jeugd. Naast veel soorten bier zijn er ook wat minder bekende bieren te vinden, zoals chocolate- en raspberrybeer. Mocht je een vervente Internetter zijn, je kunt hier gratis gebruik maken van Wi-Fi.

Annotated correctly: (10) London; 2x, 18e eeuw, Duitse; 5x, Engelse, Wi-Fi
Annotated incorrectly: (96) Mayflower; 2x, Richard, Greenwich; 2x, Union; 2x, zien, van; 5x, de; 16x, bij; 3x, pas, op; 6x, rijk, die; 2x, niet; 3x, bier; 4x, in; 5x, artikel, staat, haar, historie, plek, je; 8x, houten; 2x, restaurant, uitzicht, gok, naam, tap; 2x, voetbal, en; 6x, TV's, soorten; 2x, Thais, eten, zijn; 3x, meer, wat; 3x, kom, gratis
Not annotated: (3) The spectator, Pilgrims, Themes

TEXT 8 - TOURIST INFORMATION - Top 5: Bezienswaardigheden Krakau - reistipseuropa.nl 110511

Annotated by our dutch model

Krakau was meer dan 5 eeuwen de hoofdstad van Polen. Dat is nu niet meer het geval, maar nog steeds ademt de stad een statige en elegante sfeer uit. De stad wordt beschouwd als een van de belangrijkste toeristische trekpleistersvan Polen. Dat heeft de stad vooral te danken aan zijn architecturale en culturele rijkdom.

5. Plantypark
Begin 19e eeuw werd helaas een groot deel van de stad vernietigd, en waar vroeger de vesten zich bevonden, liggennu vaak grote ononderbroken groene parken. Deze liggen als een ring rondom het oude stadscentrum.

4. Sint Anna Kerk
De Sint Annakerk is een universiteitskerk van eind 17 eeuw. Architect van Gameren heeft het gebouw ontworpen. Kijk zeker eens rustig rond en bekijk vooral de schitterende illusionistisch fresco's en schilderijen van de hebroeders Dankwart.

3. Slowacki theater
Het Slowacki-theater, genoemd naar dichter Juliusz Slowacki, is het operagebouw van Krakau. Het gebouw stamt uit 1893 en werd geïnspireerd door de mooiste barokke opera's van Europa zoals de Opéra Garnier in Parijs. Voor debouw van de opera moest een oude kerk afgebroken worden, wat tot heel wat controverse leidde. Het operagebouwis een schitterend voorbeeld van het Poolse Eclecticisme.

2. Barbakan
Deze indrukwekkende stadspoort bevindt zich net buiten het oude stadscentrum en dateert uit de jaren 1300. Het is een van de mooiste historische

overblijfselen van het oude Europa. De poort moest elke dreiging van invallers in dekiem smoren, wat zich onder andere uit in de ruim drie meter dikke muren. Het bouwwerk bevat zeven uitkijktorens en meer dan 130 schietgaten.

1. Auschwitz
Op 60 km ten westen van Krakau, in het stadje Auschwitz (Pools: Oswiecim) ligt de stille getuige van een pijnlijkverleden: het vernietigings- en concentratiekamp Auschwitz. Tijdens de Tweede Wereldoorlog kwamen hier naar schatting tenminste 1,1 miljoen mensen, waarvan 90 % joden, om het leven. Zowel Auschwitz I, het oorspronkelijkeconcentratiekamp, als Auschwitz II, het vernietigingskamp, kan men vandaag bezichtigen.

Annotated correctly: (15) Krakau; 3x, Polen; 2x, operagebouw, Opera Garnier, Parijs, Eclecticisme, Auschwitz; 3x, Pools, vernietigings- (vernietigingskamp), Tweede Wereldoorlog
Annotated incorrectly: (0)
Not annotated: (17) Plantypark, 19e eeuw, Sint Anna Kerk; 2x, 17e eeuw, van Gameren, Dankwart, Slowacki theater; 2x, Juliusz Slowacki, barokke, Europa; 2x, Poolse, Barbakan, Auschwitz; 2x, joden

---

Annotated by the current dutch model

Krakau was meer dan 5 eeuwen de hoofdstad van Polen. Dat is nu niet meer het geval, maar nog steeds ademt de stad een statige en elegante sfeer uit. De stad wordt beschouwd als een van de belangrijkste toeristische trekpleistersvan Polen. Dat heeft de stad vooral te danken aan zijn architecturale en culturele rijkdom.

5. Plantypark
Begin 19e eeuw werd helaas een groot deel van de stad vernietigd, en waar vroeger de vesten zich bevonden, liggennu vaak grote ononderbroken groene parken. Deze liggen als een ring rondom het oude stadscentrum.

4. Sint Anna Kerk
De Sint Annakerk is een universiteitskerk van eind 17 eeuw. Architect van Gameren heeft het gebouw ontworpen. Kijk zeker eens rustig rond en bekijk vooral de schitterende illusionistisch fresco's en schilderijen van de hebroeders Dankwart.

3. Slowacki theater
Het Slowacki-theater, genoemd naar dichter Juliusz Slowacki, is het operagebouw van Krakau. Het gebouw stamt uit 1893 en werd geïnspireerd door de mooiste barokke opera's van Europa zoals de Opéra Garnier in Parijs. Voor debouw van de opera moest een oude kerk afgebroken worden, wat tot heel wat controverse leidde. Het operagebouwis een schitterend voorbeeld van het Poolse Eclecticisme.

2. Barbakan
Deze indrukwekkende stadspoort bevindt zich net buiten het oude stadscentrum en dateert uit de jaren 1300. Het is een van de mooiste historische overblijfselen van het oude Europa. De poort moest elke dreiging van invallers in dekiem smoren, wat zich onder andere uit in de ruim drie meter dikke muren. Het bouwwerk bevat zeven uitkijktorens en meer dan 130 schietgaten.

1. Auschwitz
Op 60 km ten westen van Krakau, in het stadje Auschwitz (Pools: Oswiecim) ligt de stille getuige van een pijnlijkverleden: het vernietigings- en concentratiekamp Auschwitz. Tijdens de Tweede Wereldoorlog kwamen hier naar schatting tenminste 1,1 miljoen mensen, waarvan 90 % joden, om het leven. Zowel Auschwitz I, het oorspronkelijkeconcentratiekamp, als Auschwitz II, het vernietigingskamp, kan men vandaag bezichtigen.

Annotated correctly: (15) Krakau; 3x, Polen; 4x, 19e eeuw, operagebouw; 2x, Parijs, Europa; 2x, Eclecticisme, joden
Annotated incorrectly: (73) was, meer; 3x, eeuwen; 2x, van; 15x, nu; 2x, de; 12x, stad, als; 3x, begin, groot, deel, waar, parken, ring, stadscentrum; 2x, De Sint, Architect, theater; 2x, dichter, gebouw, bouw, kerk; 2x, wat; 3x, stadspoort, net, meter, bouwwerk, km, westen, getuige, concentratiekamp; 2x, miljoen, mensen, leven, vernietigingskamp
Not annotated: (17) Plantypark, Sint Anna Kerk; 2x, 17e eeuw, van Gameren, Dankwart, Slowacki theater; 2x, Juliusz Slowacki, barokke, Opera Garnier, Barbakan, Auschwitz; 5x

---

TEXT 9 - TOURIST INFORMATION - Aanraders: Ljubljana - reistipseuropa.nl 110503

Annotated by our dutch model

Als mensen aan Slovenië denken, denken de meesten aan een arm oud Oostblokland. "Slovenië? Wat moet je daar doen?!". Toch was ik al een tijdje benieuwd naar dit land, puur vanwege een foto van een eiland met een witte kerk midden in een meer. Het bleek een kerk te zijn in Bled, een plaatsje in het noord-westen van Slovenië. Ik moest en zou hier eens naar toe gaan, en afgelopen zomer was het zover. We kampeerden in de buurt van Bled, en zijn ook een dagje naar de hoofdstad Ljubljana geweest.

Ljubljana is een kleine stad met ongeveer 310.000 inwoners en ligt centraal in het land. In principe is er, dat vind ik althans, niet heel veel te doen. Het is in ieder geval geen metropool zoals Barcelona, Parijs of Berlijn waar je een hele week zou kunnen vertoeven. Daardoor is het een ideale stad om een dag door te struinen, de bezienswaardigheden te bezoeken en een terassje te pikken. Een goed begin van een bezoek aan deze stad is het Prešerenplein. Dit plein, dat voor ons prima met de auto te bereiken was, ligt in het centrum aan de de rivier de Ljuljanica, die dwars door de stad stroomt. Het is een gezellig plein met onder andere een gebouw van de universiteit, een Franciscaanse kerk en een standbeeld van de dichter France Prešeren. Veel gebouwen aan dit plein en in de stad zijn ontworpen door de architect Jože Plecnik. Hij was onder andere verantwoordelijk voor de Franciscaanse kerk, twee bibliotheken en de drie belangrijkste bruggen van de stad.

Deze bruggen verbinden het ene deel van de stad met het andere deel, die worden gescheiden door de rivier de Ljuljanica, waarlangs je heerlijk kunt wandelen en de stad kunt verkennen. Naast de rivier staan vele kraampjes (in de zomer althans) waar de middenstand nog wat geld probeert te verdienen aan de toeristen: ideaal voor een leuk souvernir of een kaart voor thuis. Voor mij ademde de stad voornamelijk gezelligheid uit, en is het niet een stad waar je heel veel kunt doen of zien. Na een wandeling langs de rivier, door de winkelstraatjes in het centrum en over enkele mooie pleintjes heb je het meeste van de Ljubljana wel gezien. Gelukkig heeft de stad enkele mooie kerken voor de kerkliefhebber (waar ik er één van ben),

en een paar museums, zoals het Nationaal Museum.

De belangrijkste attractie van de stad is het kasteel oftewel de Ljubljanski Grad. Het middeleeuwse gebouw staat op een heuvel midden in de stad en heeft ondere andere een museum over het kasteel en over de stad, gidsen en wordt vaak gebruikt voor tentoonstellingen, bruiloften en andere speciale gelegenheden.

Ljubljana is een echte studentenstad. Hierdoor krijgt de stad een internationaal en gezellig karakter. De terrasjes zitten lekker vol en, ook een voordeel, er wordt beter Engels gesproken dan in Italië, waar we net vandaan kwamen. Ook ligt Ljubljana op korte afstand van andere bezienswaardigheden in het land, zoals de Grotten van Postojna (heel mooi en interessant, maar wel een beetje duur), de Vintgar kloof en het meer van Bled. Met de auto zou je er vanuit Nederland in één dag naar toe kunnen rijden, en het land is per vliegtuig, trein en bus ook prima bereikbaar. Zo is de stad een ideale bestemming voor een leuke groepsreis. Met de auto is de stad prima bereikbaar vanwege haar centrale ligging en er zijn voldoende goedkope parkeermogelijkheden. Hoewel de vliegticket prijsvechters de stad nog niet ontdekt hebben, gaan er dagelijks vluchten naar de hoofdstad Ljubljana vanaf Amsterdam en Brussel.

Annotated correctly: (14) Ljubljana; 6x, Barcelona, Parijs, Berlijn, France Preseren, Ljubljanski Grad, Italie, meer van Bled, Amsterdam, Brussel
Annotated incorrectly: (0)
Not annotated: (18) Slovenie; 3x, Oostblokland, Bled; 2x, Preserenplein, Ljuljanica; 2x, Franciscaanse kerk; 2x, Joze Plecnik, middeleeuwse, Engels, Postojna, Vintgar kloof, Nederland

---

Annotated by the current dutch model

Als mensen aan Slovenië denken, denken de meesten aan een arm oud Oostblokland. "Slovenië? Wat moet je daar doen?!". Toch was ik al een tijdje benieuwd naar dit land, puur vanwege een foto van een eiland met een witte kerkmidden in een meer. Het bleek een kerk te zijn in Bled, een plaatsje in het noord-westen van Slovenië. Ik moest en zou hier eens naar toe gaan, en afgelopen zomer was het zover. We kampeerden in de buurt van Bled, en zijn ook een dagje naar de hoofdstad Ljubljana geweest.

Ljubljana is een kleine stad met ongeveer 310.000 inwoners en ligt centraal in het land. In principe is er, dat vind ik althans, niet heel veel te doen. Het is in ieder geval geen metropool zoals Barcelona, Parijs of Berlijn waar je een hele week zou kunnen vertoeven. Daardoor is het een ideale stad om een dag door te struinen, de bezienswaardigheden te bezoeken en een terassje te pikken. Een goed begin van een bezoek aan deze stad is het Prešerenplein. Dit plein, dat voor ons prima met de auto te bereiken was, ligt in het centrum aan de de rivier de Ljuljanica, die dwars door de stad stroomt. Het is een gezellig plein met onder andere een gebouw van de universiteit, een Franciscaanse kerk en een standbeeld van de dichter France Prešeren. Veel gebouwen aan dit plein en in de stad zijn ontworpen door dearchitect Jože Plecnik. Hij was onder andere verantwoordelijk voor de Franciscaanse kerk, twee bibliotheken en dedrie belangrijkste bruggen van de stad.

Deze bruggen verbinden het ene deel van de stad met het andere deel, die worden gescheiden door de rivier deLjuljanica, waarlangs je heerlijk kunt wandelen en de stad kunt verkennen. Naast de rivier staan vele kraampjes (in dezomer althans) waar de middenstand nog wat geld probeert te verdienen aan de toeristen: ideaal voor een leuk souvenir of een kaart voor thuis. Voor mij ademde de stad voornamelijk gezelligheid uit, en is het niet een stad waar je heel veel kunt doen of zien. Na een wandeling langs de rivier, door de winkelstraatjes in het centrum en over enkele mooie pleintjes heb je het meeste van de Ljubljana wel gezien. Gelukkig heeft de stad enkele mooie kerken voor dekerkliefhebber (waar ik er één van ben), en een paar museums, zoals het Nationaal Museum.

De belangrijkste attractie van de stad is het kasteel oftewel de Ljubljanski Grad. Het middeleeuwse gebouw staat opeen heuvel midden in de stad en heeft ondere andere een museum over het kasteel en over de stad, gidsen en wordt vaak gebruikt voor tentoonstellingen, bruiloften en andere speciale gelegenheden.

Ljubljana is een echte studentenstad. Hierdoor krijgt de stad een internationaal en gezellig karakter. De terrasjes zitten lekker vol en, ook een voordeel, er wordt beter Engels gesproken dan in Italië, waar we net vandaan kwamen. Ook ligt Ljubljana op korte afstand van andere bezienswaardigheden in het land, zoals de Grotten van Postojna (heel mooi en interessant, maar wel een beetje duur), de Vintgar kloof en het meer van Bled. Met de auto zou je er vanuit Nederland in één dag naar toe kunnen rijden, en het land is per vliegtuig, trein en bus ook prima bereikbaar. Zo is de stad een ideale bestemming voor een leuke groepsreis. Met de auto is de stad prima bereikbaar vanwege haar centrale ligging en er zijn voldoende goedkope parkeermogelijkheden. Hoewel de vliegticket prijsvechters de stad nog niet ontdekt hebben, gaan er dagelijks vluchten naar de hoofdstad Ljubljana vanaf Amsterdam en Brussel.

Annotated correctly: (12) Oostblokland, Ljubljana; 4x, Parijs, Berlijn, middeleeuwse, Engels, Nederland, Amsterdam, Brussel
Annotated incorrectly: (83) mensen, de; 29x, land, kerk, in; 15x, zomer; 2x, stad; 3x, Barcelona (City, links to FC Barcelona), dag; 2x, auto; 3x, centrum; 2x, rivier; 4x, universiteit, Franciscaanse kerk; 2x, stadbeeld, dichter, architect, geld, toeristen, kasteel; 2x, staat, op; 2x, museum, karakter, net, vliegtuig, trein, bus
Not annotated: (14) Slovenie; 3x, Bled; 2x, Preserenplein, Ljuljanica; 2x, Joze Plecnik, Ljubljanski Grad, Italie, Postojna, Vintgar kloof, meer van Bled

---