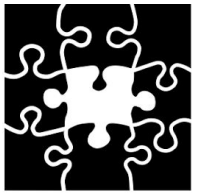
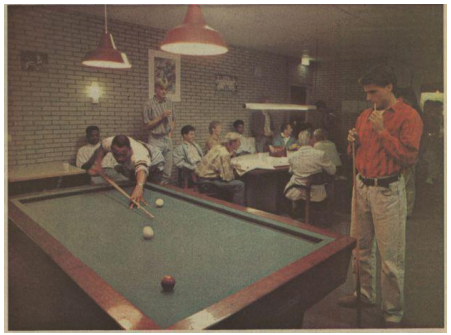
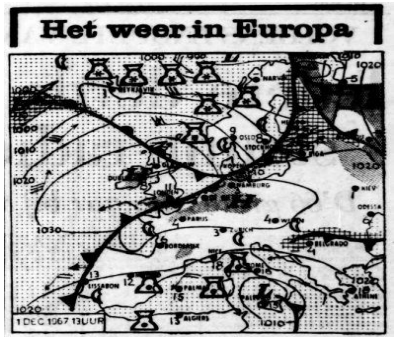
1 Million Dutch Captioned Newspaper Images

Desmond Elliott and Martijn Kleppe



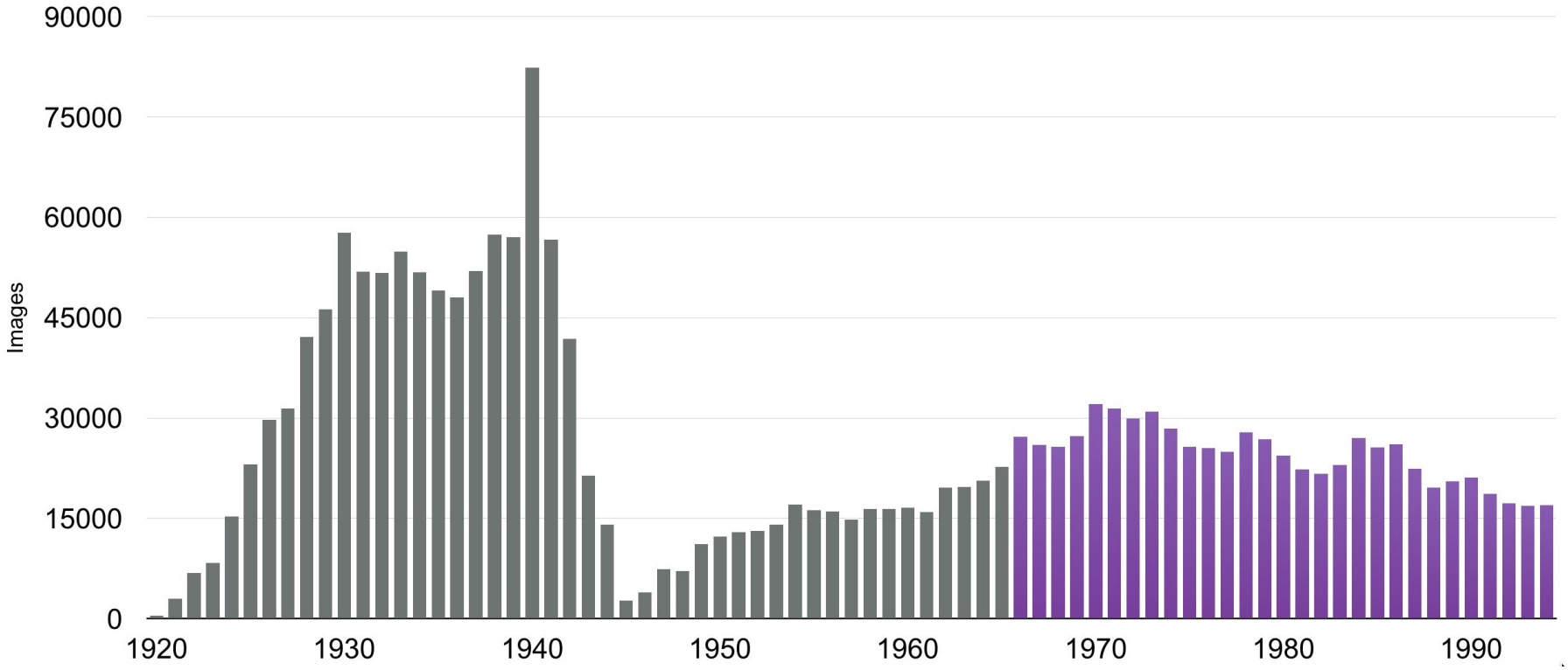
The KBK-1M Dataset



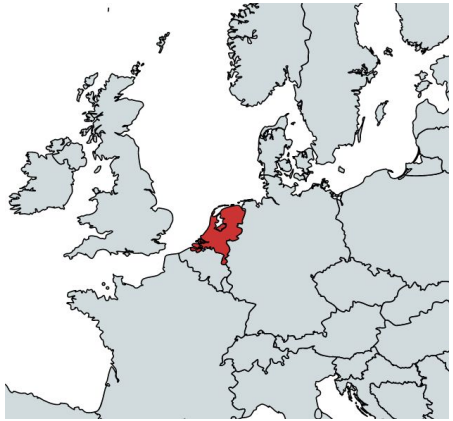


- KBK-1M is 1.6 million images in newspaper context
- Captions via Optical Character Recognition
- **K**oninklijke **B**ibliothee**k** archive from 1922 - 1994

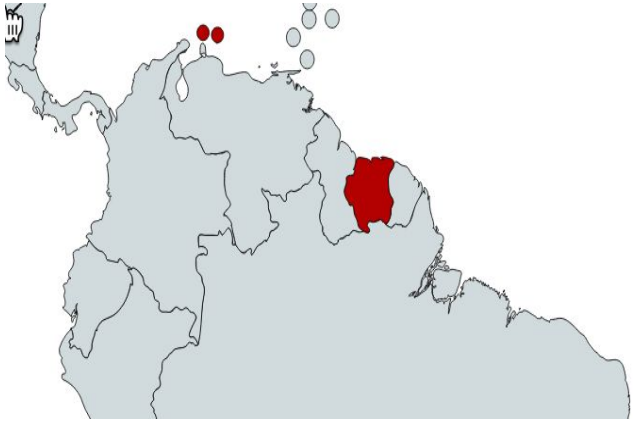
A 20th Century Corpus



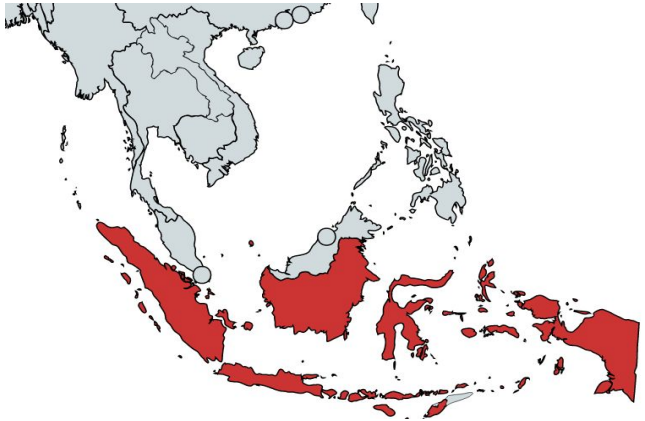
Across the World



Netherlands

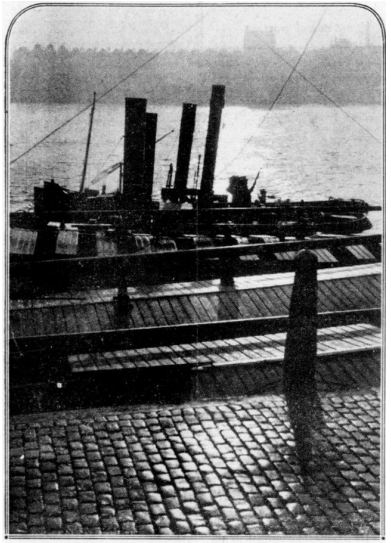


Suriname,
Curaçao,
Dutch Antillies

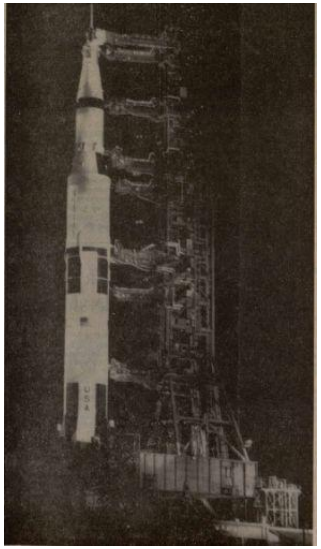


Indonesia

The World Through (Dutch) Time



De Telegraaf, 30-11-1935



Nederlands Dagblad, 15-07-1969

Amigoe di Curacao, 13-11-1989

1920

1990

Collecting the Dataset

De Niewsgier, 27-01-1951

Collecting the Dataset


NIEUWSGIER

Natsir voor parlement

Heden regeringsverklaring over Assaats ontslagverzoek

VN-troepen hernemen Suwon

De Burmese minister van buitenlandse zaken vertoeft te Djakarta met enige leden van zijn staf; hij werd op Kemajoran verwelkomd door Indonesische autoriteiten en de Burmese ambassadeur. (Ipphos) j

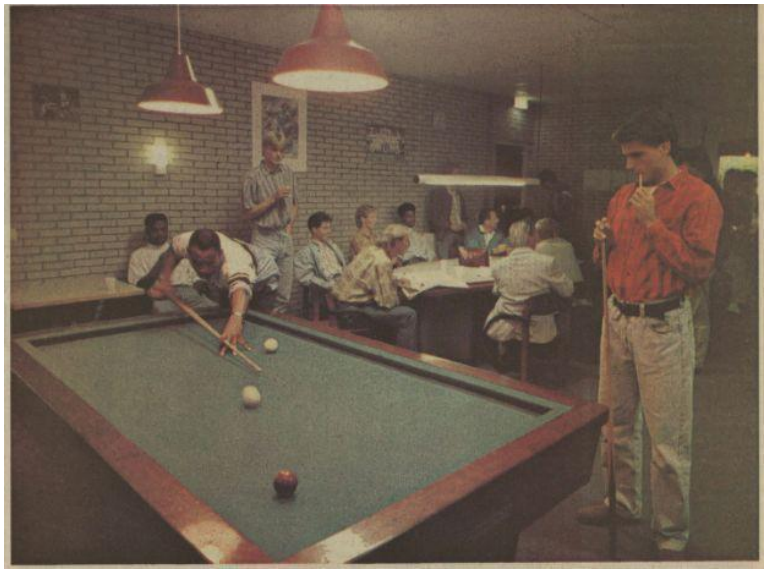
De Niewsgier, 27-01-1951

People

20 DAGEN IN CEL WASHINGTON — Dominee Ralph Abernathy, die maandag werd angehouden toen hij een mars wilde organiseren van de „stad van de Wederopstanding naar het kapitool, is dinsdag wegens het organiseren van een mars zonder toestemming op het terrein van het kapitool tot 20 dagen hechtenis veroordeeld.

(Friese koerier: 26-06-1968)

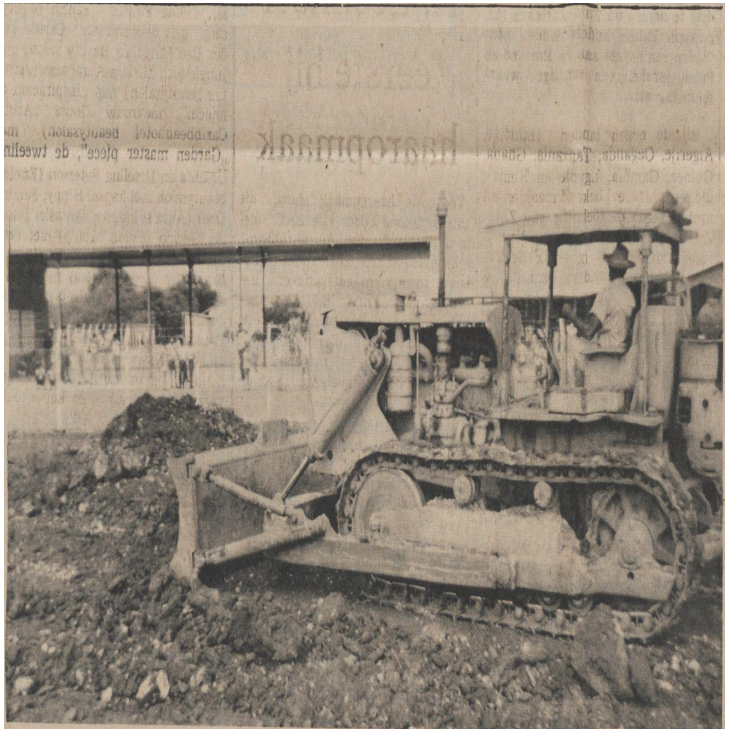
People in Places



Biljarten in plaats van voetballen. Mart van Duren (rechts) kijk naar een poking van Henri Meijer

(Nieuwsblad van het Noorden: 16-09-1991)

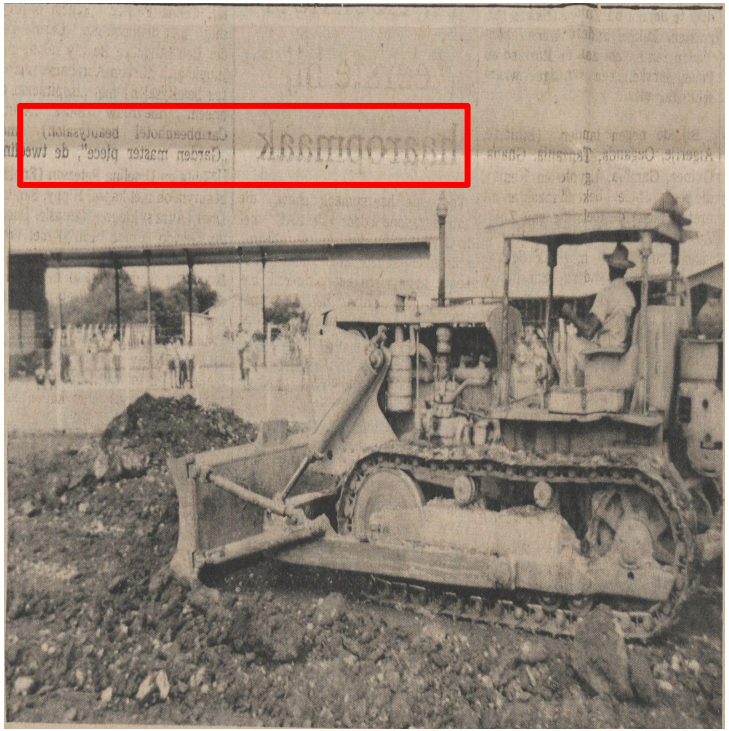
Places



De graafwerkzaamheden voor -de bouw van vier school lokalen, toegezegd door de heer H. Rochcreau van het EEG-ontwikkelingsfoads tijdens diens bezoek aan Bonaire, zijn reeds begonnen.

(Amigoe di Curacao : weekblad voor de Curacaosche eilanden, 19-02-1968

Places

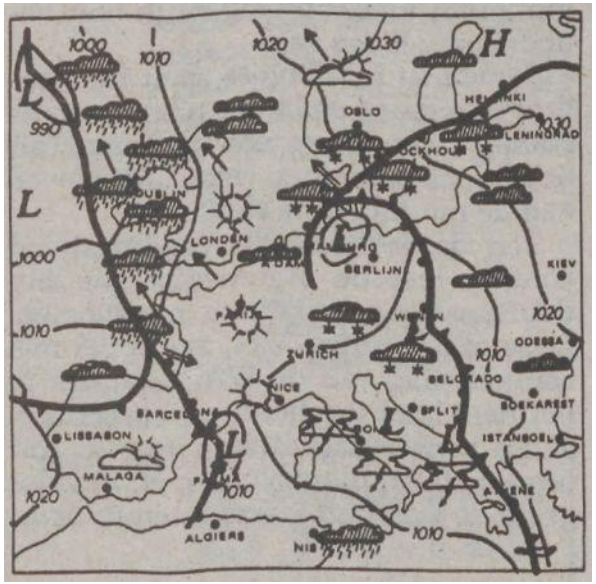


De graafwerkzaamheden voor -de bouw van vier school lokalen, toegezegd door de heer H. Rochcreau van het EEG-ontwikkelingsfoads tijdens diens bezoek aan Bonaire, zijn reeds begonnen.

(Amigoe di Curacao : weekblad voor de Curacaosche eilanden, 19-02-1968

Weather Maps

(De Tijd, 02-12-1967)



(De Waarheid, 05-04-1984)

Opklaringen

Aan de periode met lage temperaturen komt geleidelijk een einde. Er waait geen koude noordelijke wind meer [...]

Comic strips

(De Telegraaf)

KBK-1M compared to other datasets

	Images	Notes
Newspapers		
KBK-1M	1,603,396	Dutch newspapers
ION (Hollink et al., 2016)	300,000	Online News
BBC News (Feng and Lapata, 2008)	3,361	BBC News Online

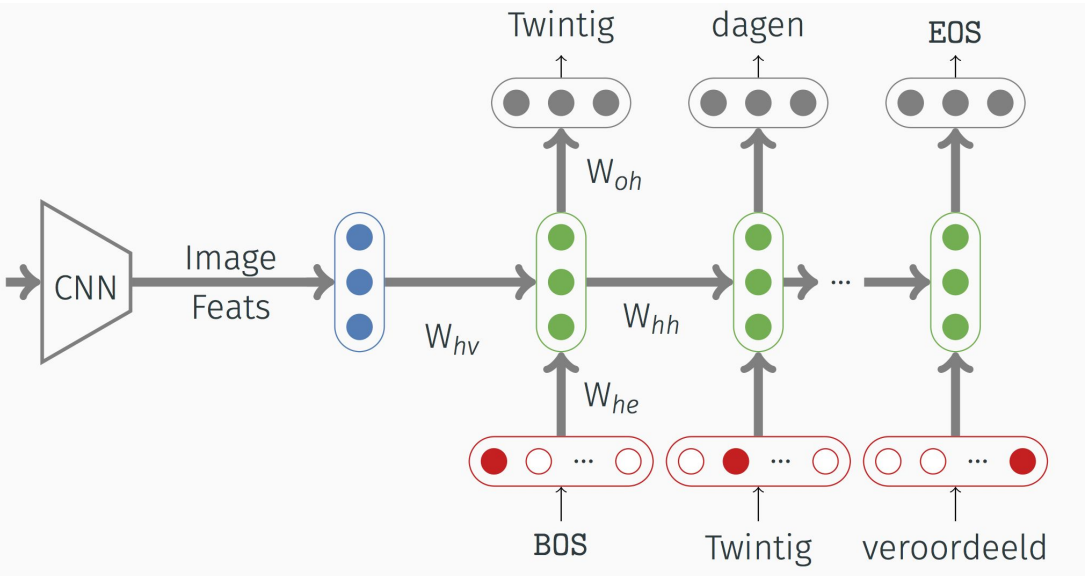
KBK-1M compared to other datasets

	Images	Notes
Newspapers		
KBK-1M	1,603,396	Dutch newspapers
ION (Hollink et al., 2016)	300,000	Online News
BBC News (Feng and Lapata, 2008)	3,361	BBC News Online
User captions		
SBU 1M (Ordonez et al., 2011)	1,000,000	–
Deja-Images (Chen et al., 2015a)	440,000	–

KBK-1M compared to other datasets

	Images	Notes
Newspapers		
KBK-1M	1,603,396	Dutch newspapers
ION (Hollink et al., 2016)	300,000	Online News
BBC News (Feng and Lapata, 2008)	3,361	BBC News Online
User captions		
SBU 1M (Ordonez et al., 2011)	1,000,000	–
Deja-Images (Chen et al., 2015a)	440,000	–
Crowdsourced descriptions		
Flickr30K (Hodosh et al., 2013b)	30,000	–
Multi30K (Elliott et al., 2016)	30,000	Bilingual text
MS COCO (Chen et al., 2015b)	164,062	–

KBK-1M for Language Generation



Vinyals et al (CVPR 2015); Karpathy and Fei-Fei (CVPR 2015); *inter-alia*.

KBK-1M for Multimodal Weather Report Generation

Temperature			
Time	Min	Mean	Max
06:00-21:00	9	15	21

Cloud Sky Cover	
Time	Percent (%)
06:00-09:00	25-50
09:00-12:00	50-75

Wind Speed			
Time	Min	Mean	Max
06:00-21:00	15	20	30

Wind Direction	
Time	Mode
06:00-21:00	S

Cloudy, with a low around 10. South wind around 20 mph.

Konstas and Lapata (ACL 2012)

KBK-1M for Multimodal Weather Report Generation

Temperature			
Time	Min	Mean	Max
06:00-21:00	9	15	21

Cloud Sky Cover	
Time	Percent (%)
06:00-09:00	25-50
09:00-12:00	50-75

Wind Speed			
Time	Min	Mean	Max
06:00-21:00	15	20	30

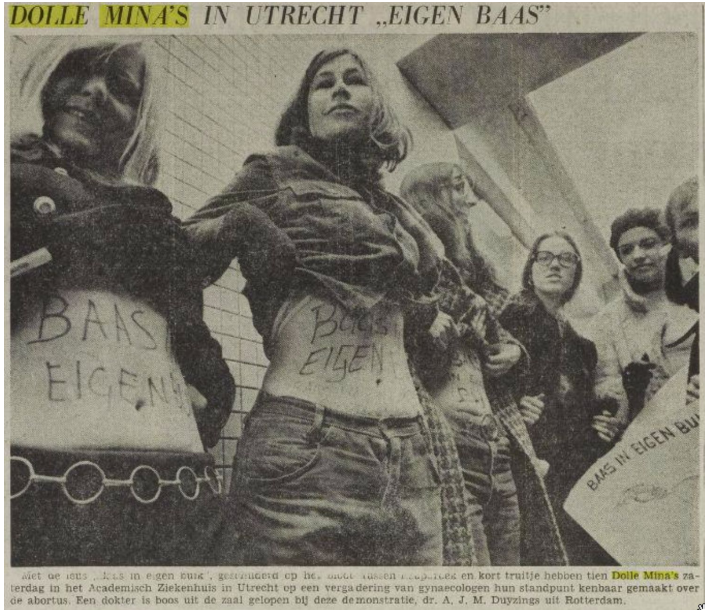
Wind Direction	
Time	Mode
06:00-21:00	S

Cloudy, with a low around 10. South wind around 20 mph.

Konstas and Lapata (ACL 2012)

KBK-1M for Digital Humanities Research

- How can we automatically find reused images over time?



Kleppe (2013)

Summary

- KBK-1M is 1.6 million images in Dutch newspapers
- Spanning three continents from 1922 - 1994
- Captioned images via Optical Character Recognition
- Available as year-by-year archives (517GB total)
- Get it at http://lab.kbresearch.nl/

What happened in the 1920s?

- The emergence of “photo pages” in the central sheets of newspapers resulted in an explosion of photographic use in newspapers.