# Semester Project – Jan Linder

Overview week 6

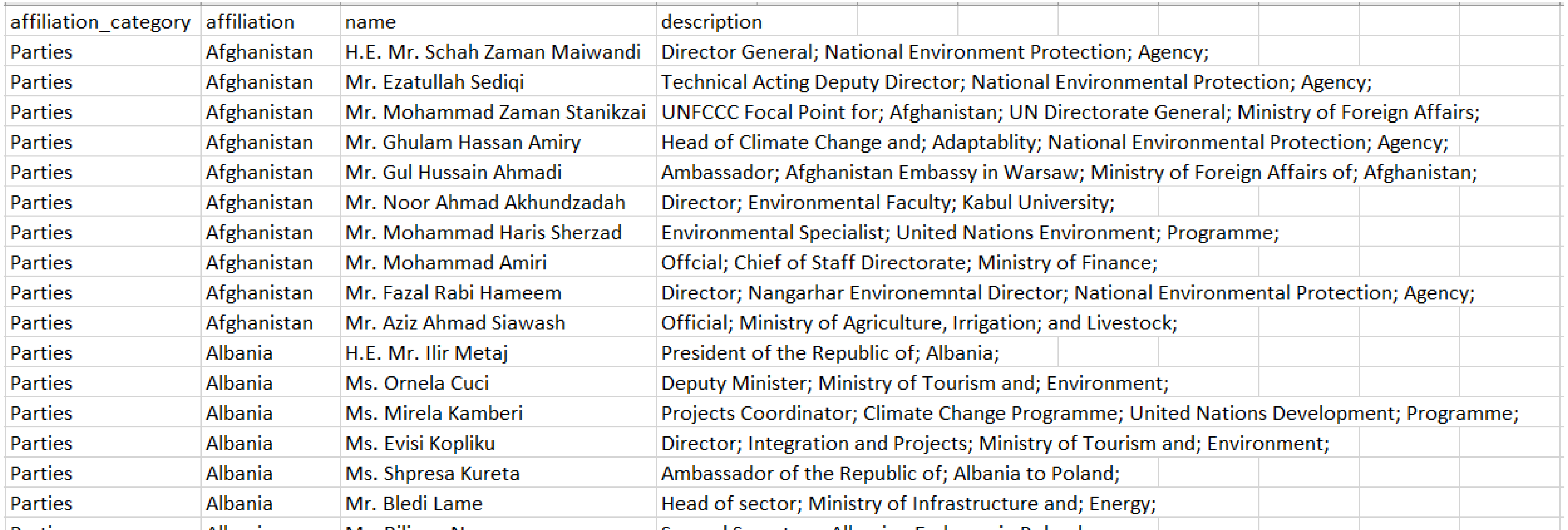
23.10.2020

# Part I: Participant lists

- In a first step, the PDFs needed to be transformed to .txt files
  - Scans: Used the open-source tool *tesseract* that is developed by Google since 2006. It performs Optical Character Recognition (OCR) on the lists.
  - “Normal” PDFs: Used the open-source tool *pdfminer* with minor changes that allow to extract several columns in the right order.

# Part I: Participant lists

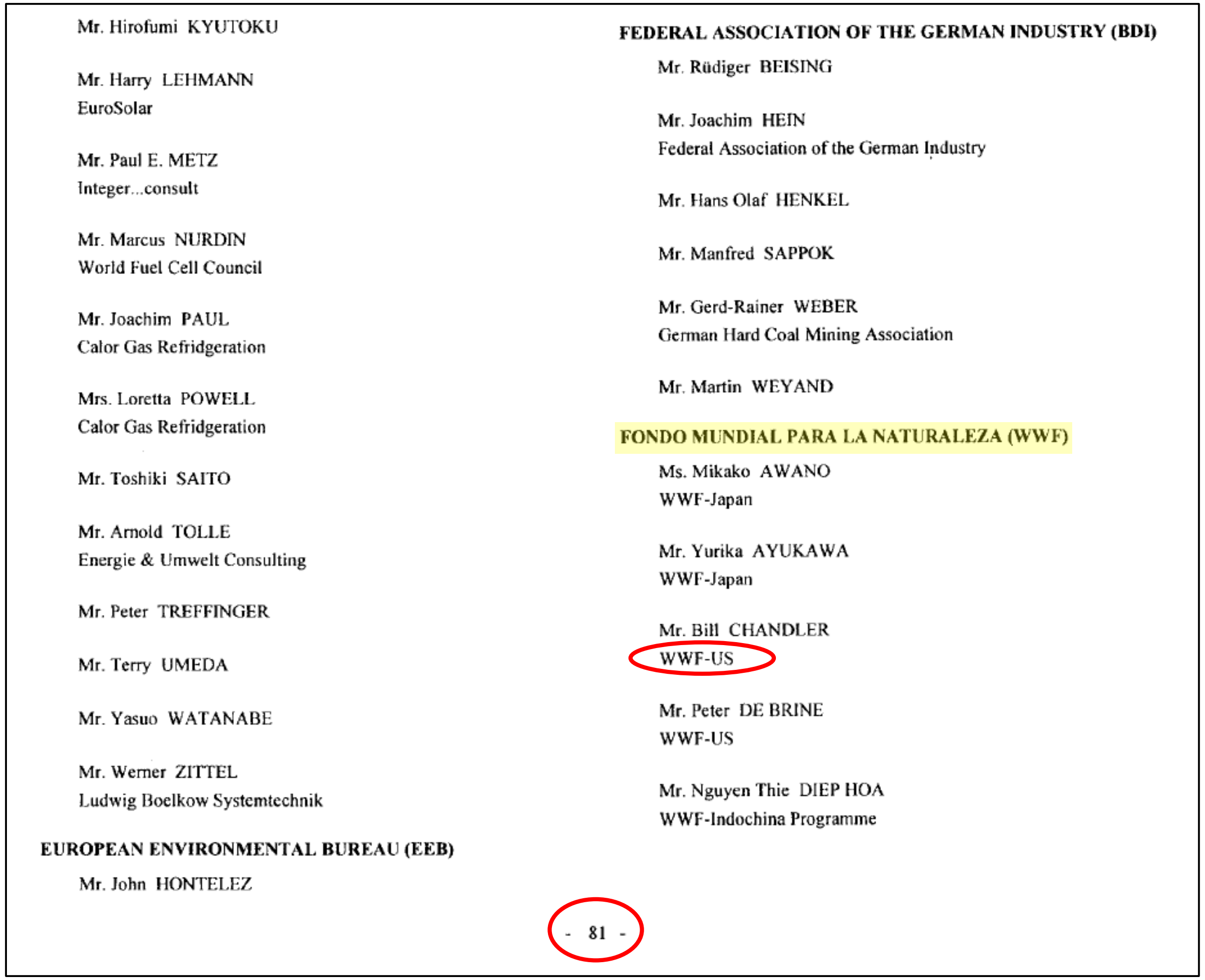
- In a first step, the PDFs needed to be transformed to .txt files
- Then, extract a data structure containing the participants from the .txt files: *affiliation category, affiliation, name, description*



| affiliation_category | affiliation | name | description |
|---|---|---|---|
| Parties | Afghanistan | H.E. Mr. Schah Zaman Maiwandi | Director General; National Environment Protection; Agency; |
| Parties | Afghanistan | Mr. Ezatullah Sediqi | Technical Acting Deputy Director; National Environmental Protection; Agency; |
| Parties | Afghanistan | Mr. Mohammad Zaman Stanikzai | UNFCCC Focal Point for; Afghanistan; UN Directorate General; Ministry of Foreign Affairs; |
| Parties | Afghanistan | Mr. Ghulam Hassan Amiry | Head of Climate Change and; Adaptablity; National Environmental Protection; Agency; |
| Parties | Afghanistan | Mr. Gul Hussain Ahmadi | Ambassador; Afghanistan Embassy in Warsaw; Ministry of Foreign Affairs of; Afghanistan; |
| Parties | Afghanistan | Mr. Noor Ahmad Akhundzadah | Director; Environmental Faculty; Kabul University; |
| Parties | Afghanistan | Mr. Mohammad Haris Sherzad | Environmental Specialist; United Nations Environment; Programme; |
| Parties | Afghanistan | Mr. Mohammad Amiri | Offcial; Chief of Staff Directorate; Ministry of Finance; |
| Parties | Afghanistan | Mr. Fazal Rabi Hameem | Director; Nangarhar Environemntal Director; National Environmental Protection; Agency; |
| Parties | Afghanistan | Mr. Aziz Ahmad Siawash | Official; Ministry of Agriculture, Irrigation; and Livestock; |
| Parties | Albania | H.E. Mr. Ilir Metaj | President of the Republic of; Albania; |
| Parties | Albania | Ms. Ornela Cuci | Deputy Minister; Ministry of Tourism and; Environment; |
| Parties | Albania | Ms. Mirela Kamberi | Projects Coordinator; Climate Change Programme; United Nations Development; Programme; |
| Parties | Albania | Ms. Evisi Kopliku | Director; Integration and Projects; Ministry of Tourism and; Environment; |
| Parties | Albania | Ms. Shpresa Kureta | Ambassador of the Republic of; Albania to Poland; |
| Parties | Albania | Mr. Bledi Lame | Head of sector; Ministry of Infrastructure and; Energy; |

# Example: Scans

## COP 3



Mr. Hirofumi KYUTOKU

Mr. Harry LEHMANN
EuroSolar

Mr. Paul E. METZ
Integer...consult

Mr. Marcus NURDIN
World Fuel Cell Council

Mr. Joachim PAUL
Calor Gas Refridgeration

Mrs. Loretta POWELL
Calor Gas Refridgeration

Mr. Toshiki SAITO

Mr. Arnold TOLLE
Energie & Umwelt Consulting

Mr. Peter TREFFINGER

Mr. Terry UMEDA

Mr. Yasuo WATANABE

Mr. Werner ZITTEL
Ludwig Boelkow Systemtechnik

**EUROPEAN ENVIRONMENTAL BUREAU (EEB)**

Mr. John HONTELEZ

**FEDERAL ASSOCIATION OF THE GERMAN INDUSTRY (BDI)**

Mr. Rüdiger BEISING

Mr. Joachim HEIN
Federal Association of the German Industry

Mr. Hans Olaf HENKEL

Mr. Manfred SAPPOK

Mr. Gerd-Rainer WEBER
German Hard Coal Mining Association

Mr. Martin WEYAND

**FONDO MUNDIAL PARA LA NATURALEZA (WWF)**

Ms. Mikako AWANO
WWF-Japan

Mr. Yurika AYUKAWA
WWF-Japan

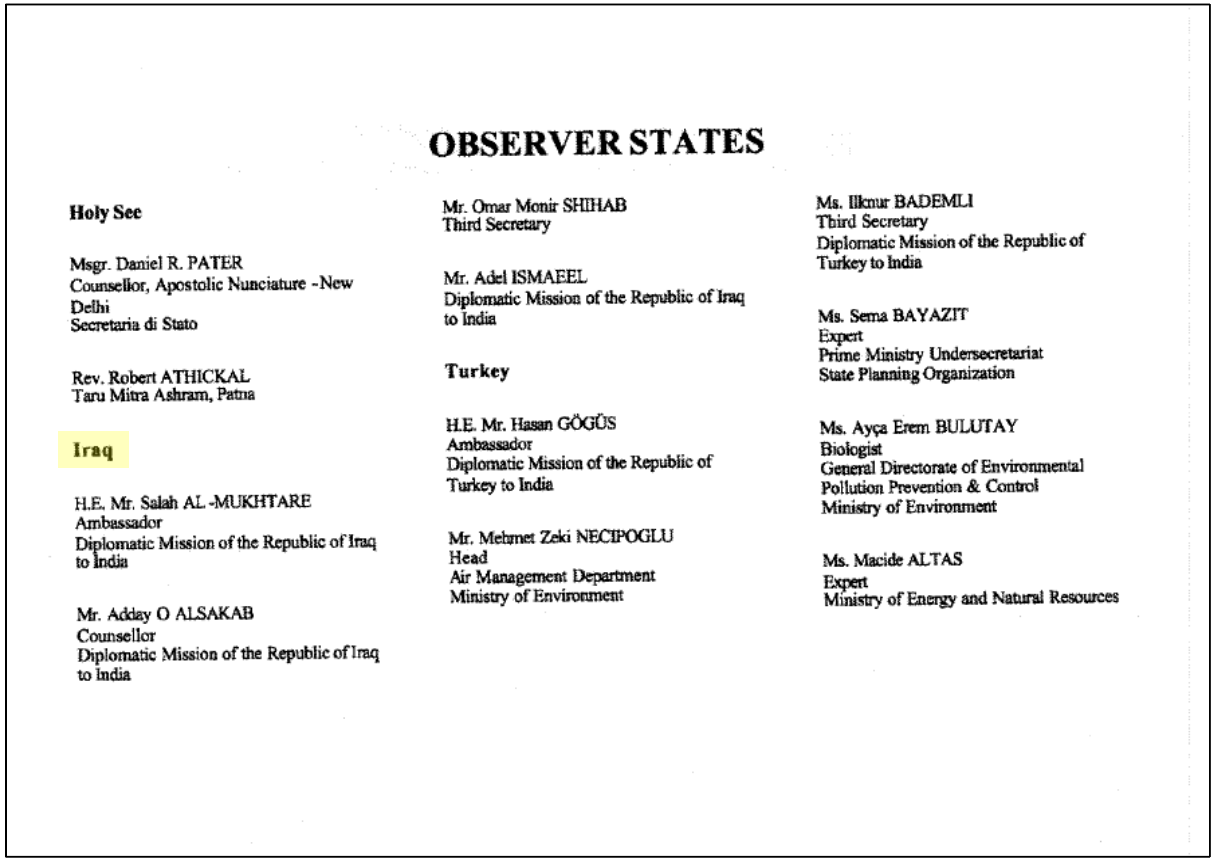
Mr. Bill CHANDLER
WWF-US

Mr. Peter DE BRINE
WWF-US

Mr. Nguyen Thie DIEP HOA
WWF-Indochina Programme

- 81 -

## COP 8



**OBSERVER STATES**

**Holy See**

Msgr. Daniel R. PATER
Counsellor, Apostolic Nunciature -New Delhi
Secretaria di Stato

Rev. Robert ATHICKAL
Taru Mitra Ashram, Patna

**Iraq**

H.E. Mr. Salah AL -MUKHTARE
Ambassador
Diplomatic Mission of the Republic of Iraq to India

Mr. Adday O ALSAKAB
Counsellor
Diplomatic Mission of the Republic of Iraq to India

Mr. Omar Monir SHIHAB
Third Secretary

Mr. Adel ISMAEEL
Diplomatic Mission of the Republic of Iraq to India

**Turkey**

H.E. Mr. Hasan GÖGÜS
Ambassador
Diplomatic Mission of the Republic of Turkey to India

Mr. Mehmet Zeki NECIPOGLU
Head
Air Management Department
Ministry of Environment

Ms. Ilknur BADEMLI
Third Secretary
Diplomatic Mission of the Republic of Turkey to India

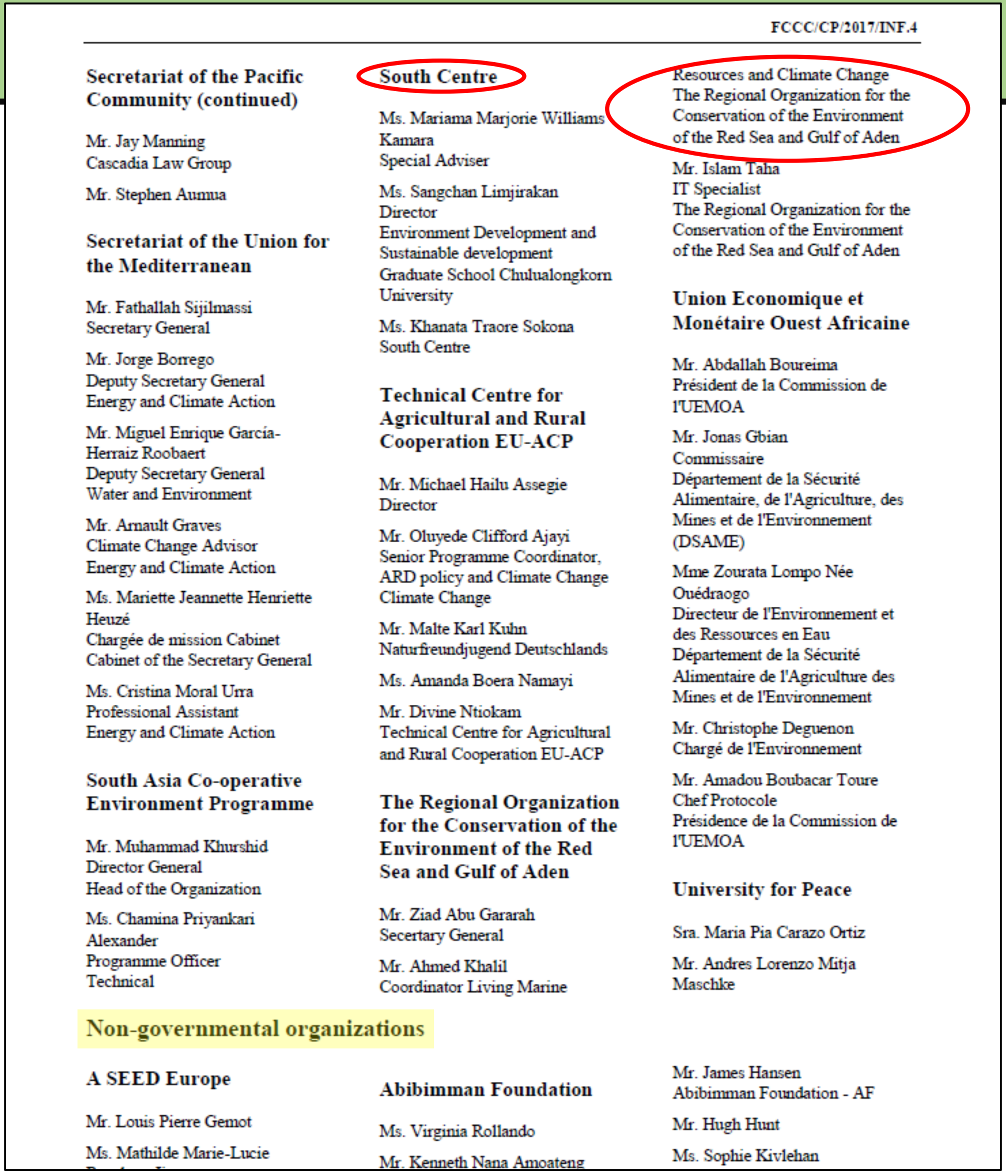
Ms. Sema BAYAZIT
Expert
Prime Ministry Undersecretariat
State Planning Organization

Ms. Ayça Erem BULUTAY
Biologist
General Directorate of Environmental Pollution Prevention & Control
Ministry of Environment

Ms. Macide ALTAS
Expert
Ministry of Energy and Natural Resources

# Example: newer PDFs

COP 23



FCCC/CP/2017/INF.4

### Secretariat of the Pacific Community (continued)

Mr. Jay Manning
Cascadia Law Group

Mr. Stephen Aumua

### Secretariat of the Union for the Mediterranean

Mr. Fathallah Sijilmassi
Secretary General

Mr. Jorge Borrego
Deputy Secretary General
Energy and Climate Action

Mr. Miguel Enrique García-Herraiz Roobaert
Deputy Secretary General
Water and Environment

Mr. Arnault Graves
Climate Change Advisor
Energy and Climate Action

Ms. Mariette Jeannette Henriette Heuzé
Chargée de mission Cabinet
Cabinet of the Secretary General

Ms. Cristina Moral Urra
Professional Assistant
Energy and Climate Action

### South Asia Co-operative Environment Programme

Mr. Muhammad Khurshid
Director General
Head of the Organization

Ms. Chamina Priyankari Alexander
Programme Officer
Technical

### South Centre

Ms. Mariama Marjorie Williams Kamara
Special Adviser

Ms. Sangchan Limjirakan
Director
Environment Development and Sustainable development
Graduate School Chulualongkorn University

Ms. Khanata Traore Sokona
South Centre

### Technical Centre for Agricultural and Rural Cooperation EU-ACP

Mr. Michael Hailu Assegie
Director

Mr. Oluyede Clifford Ajayi
Senior Programme Coordinator, ARD policy and Climate Change
Climate Change

Mr. Malte Karl Kuhn
Naturfreundjugend Deutschlands

Ms. Amanda Boera Namayi

Mr. Divine Ntiokam
Technical Centre for Agricultural and Rural Cooperation EU-ACP

### The Regional Organization for the Conservation of the Environment of the Red Sea and Gulf of Aden

Mr. Ziad Abu Gararah
Secertary General

Mr. Ahmed Khalil
Coordinator Living Marine
Resources and Climate Change
The Regional Organization for the Conservation of the Environment of the Red Sea and Gulf of Aden

Mr. Islam Taha
IT Specialist
The Regional Organization for the Conservation of the Environment of the Red Sea and Gulf of Aden

### Union Economique et Monétaire Ouest Africaine

Mr. Abdallah Boureima
Président de la Commission de l'UEMOA

Mr. Jonas Gbian
Commissaire
Département de la Sécurité Alimentaire, de l'Agriculture, des Mines et de l'Environnement (DSAME)

Mme Zourata Lompo Née Ouédraogo
Directeur de l'Environnement et des Ressources en Eau
Département de la Sécurité Alimentaire de l'Agriculture des Mines et de l'Environnement

Mr. Christophe Deguenon
Chargé de l'Environnement

Mr. Amadou Boubacar Toure
Chef Protocole
Présidence de la Commission de l'UEMOA

### University for Peace

Sra. Maria Pia Carazo Ortiz

Mr. Andres Lorenzo Mitja Maschke

## Non-governmental organizations

### A SEED Europe

Mr. Louis Pierre Gemot

Ms. Mathilde Marie-Lucie

### Abibimman Foundation

Ms. Virginia Rollando

Mr. Kenneth Nana Amoateng

Mr. James Hansen
Abibimman Foundation - AF

Mr. Hugh Hunt

Ms. Sophie Kivlehan

# Part I: Participant lists

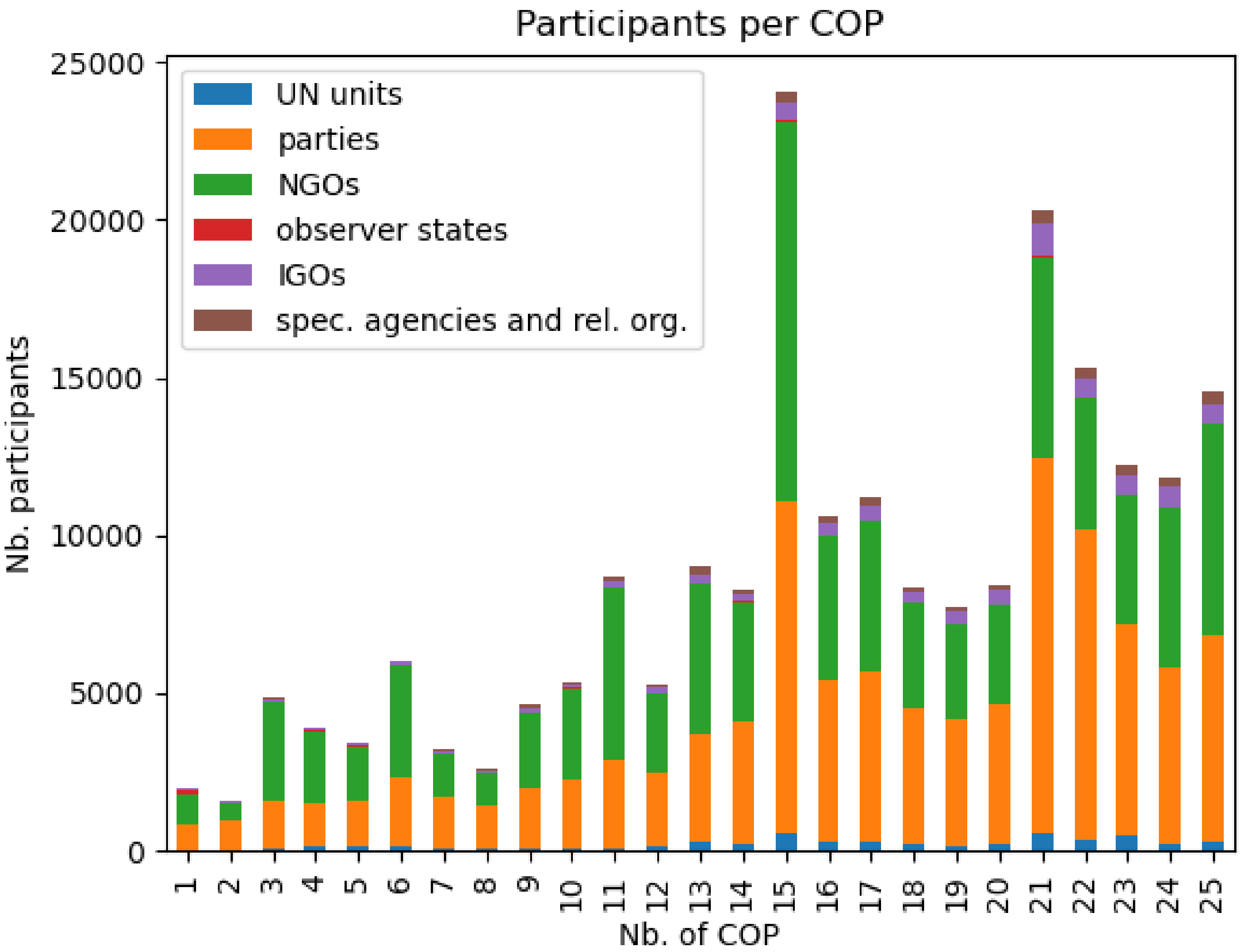
Challenges:

- Design inconsistencies in the lists lead to minor errors
  - Missed participants
  - Wrong affiliation or affiliation category
  - Typos (especially for OCR)
  - Long names might not be detected entirely
- Corrigenda are not (yet) considered

# Intermediate results

Extracted participants per COP
The colours indicate the affiliation category:

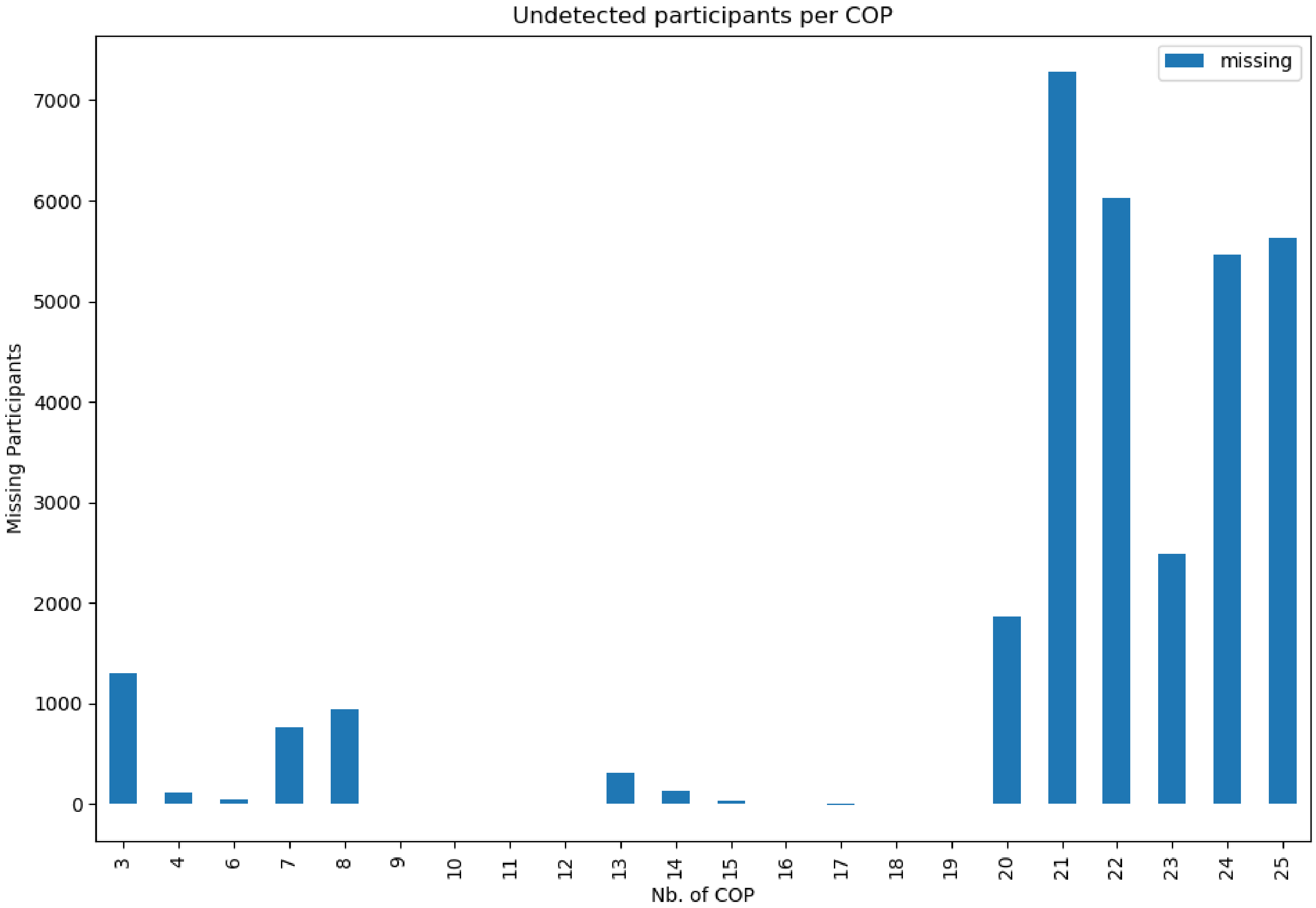
- Parties
- Observer States
- United Nations secretariat units and bodies
- Specialized agencies and related organizations
- Intergovernmental organizations
- Non-governmental organizations



# Intermediate results

Undetected participants
Comparison of my extracted number of participants and the stated number on the first page of the participant list.

We expect the discrepancy (at least for the newer lists) to be mainly due to the source, as Victor got similar results for COP24 and COP 25 with a totally different method.



# Appendix: How it works – tesseract

Find connected components (nested)

1. Outlines of all elements are grouped to blobs
2. Blobs are organized into text lines
3. Text lines are broken into words

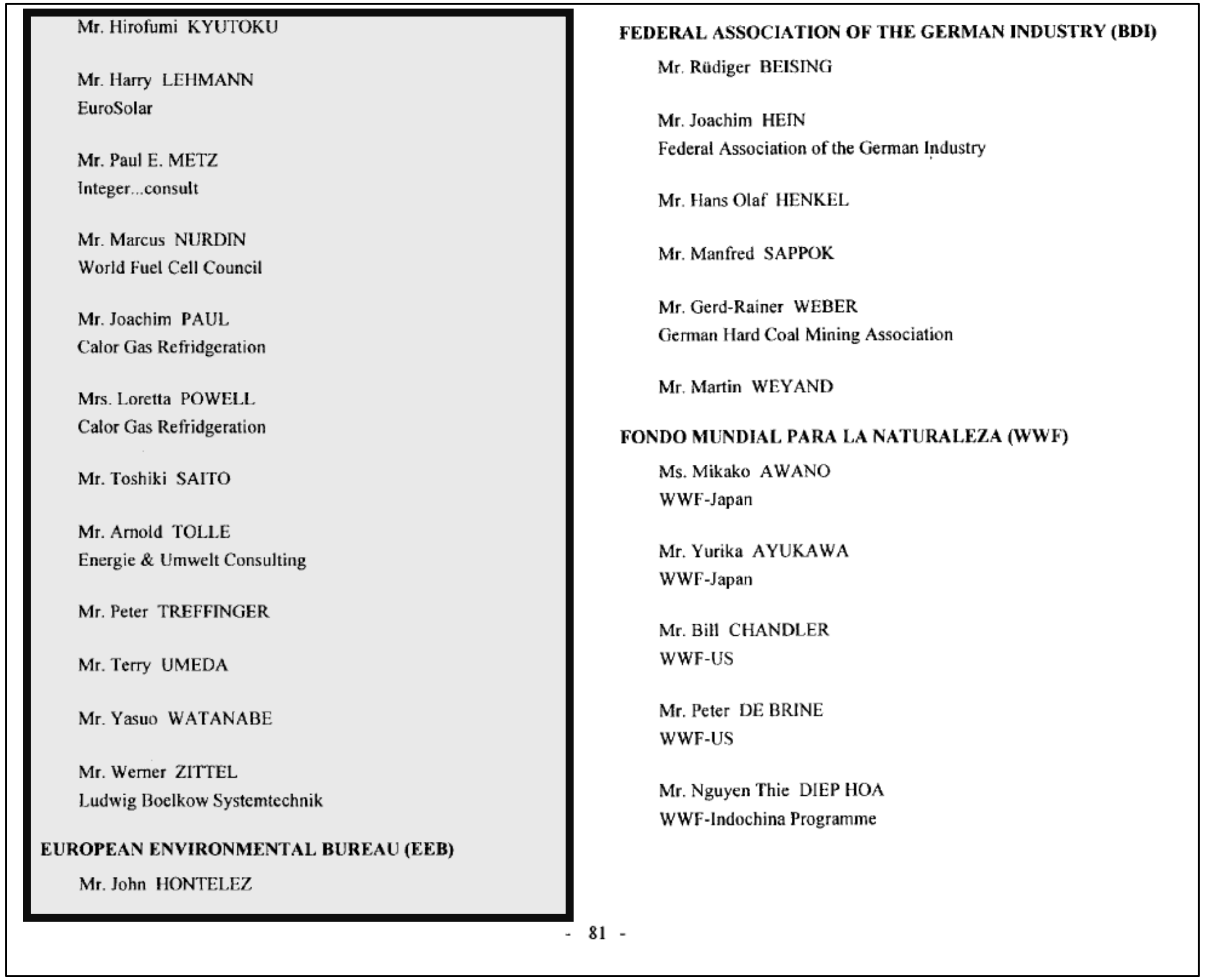
Recognition: Two-pass process

1. Guess each word (also with dictionary) and “learn” from satisfactory results
2. Go through everything again using all the collected information

Explanation inspired by: https://www.quora.com/How-does-the-Tesseract-API-for-OCR-work?share=1

# Appendix: How it works - tesseract

To prevent the two columns from being mixed I inserted a box on certain pages. This helps to find accurate blobs.



Mr. Hirofumi KYUTOKU

Mr. Harry LEHMANN
EuroSolar

Mr. Paul E. METZ
Integer...consult

Mr. Marcus NURDIN
World Fuel Cell Council

Mr. Joachim PAUL
Calor Gas Refridgeration

Mrs. Loretta POWELL
Calor Gas Refridgeration

Mr. Toshiki SAITO

Mr. Arnold TOLLE
Energie & Umwelt Consulting

Mr. Peter TREFFINGER

Mr. Terry UMEDA

Mr. Yasuo WATANABE

Mr. Werner ZITTEL
Ludwig Boelkow Systemtechnik

**EUROPEAN ENVIRONMENTAL BUREAU (EEB)**

Mr. John HONTELEZ

**FEDERAL ASSOCIATION OF THE GERMAN INDUSTRY (BDI)**

Mr. Rüdiger BEISING

Mr. Joachim HEIN
Federal Association of the German Industry

Mr. Hans Olaf HENKEL

Mr. Manfred SAPPOK

Mr. Gerd-Rainer WEBER
German Hard Coal Mining Association

Mr. Martin WEYAND

**FONDO MUNDIAL PARA LA NATURALEZA (WWF)**

Ms. Mikako AWANO
WWF-Japan

Mr. Yurika AYUKAWA
WWF-Japan

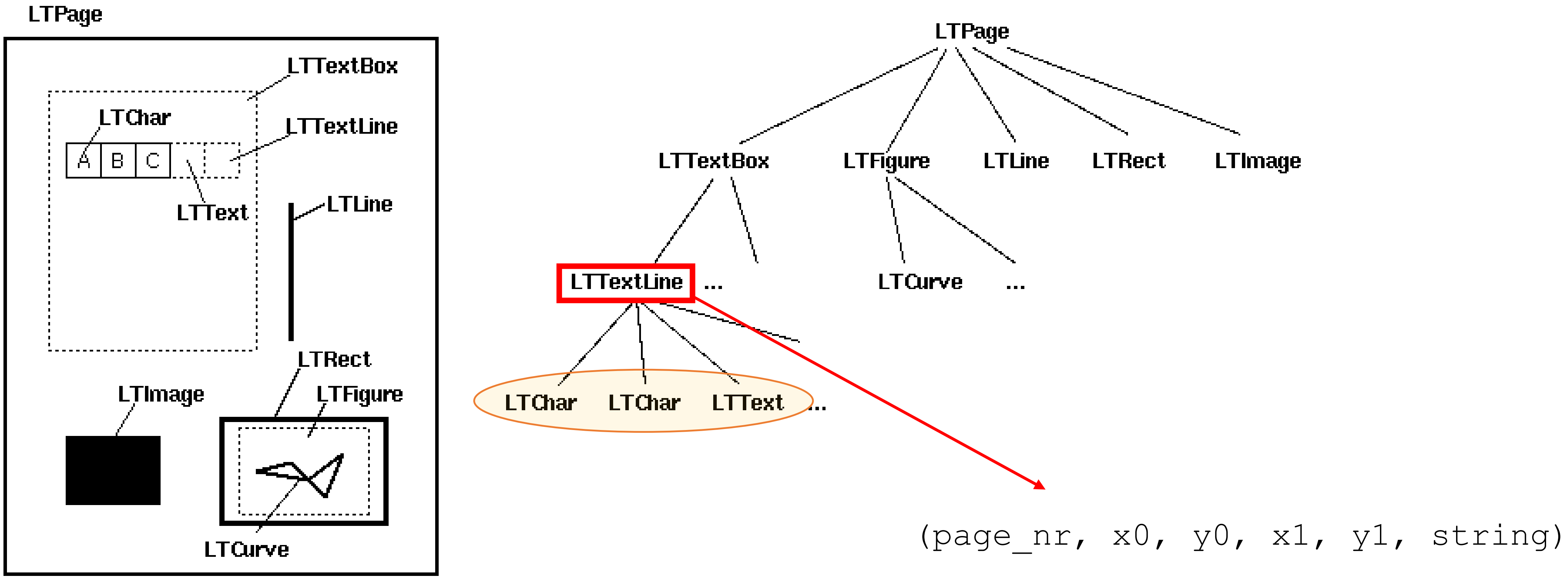
Mr. Bill CHANDLER
WWF-US

Mr. Peter DE BRINE
WWF-US

Mr. Nguyen Thie DIEP HOA
WWF-Indochina Programme

- 81 -

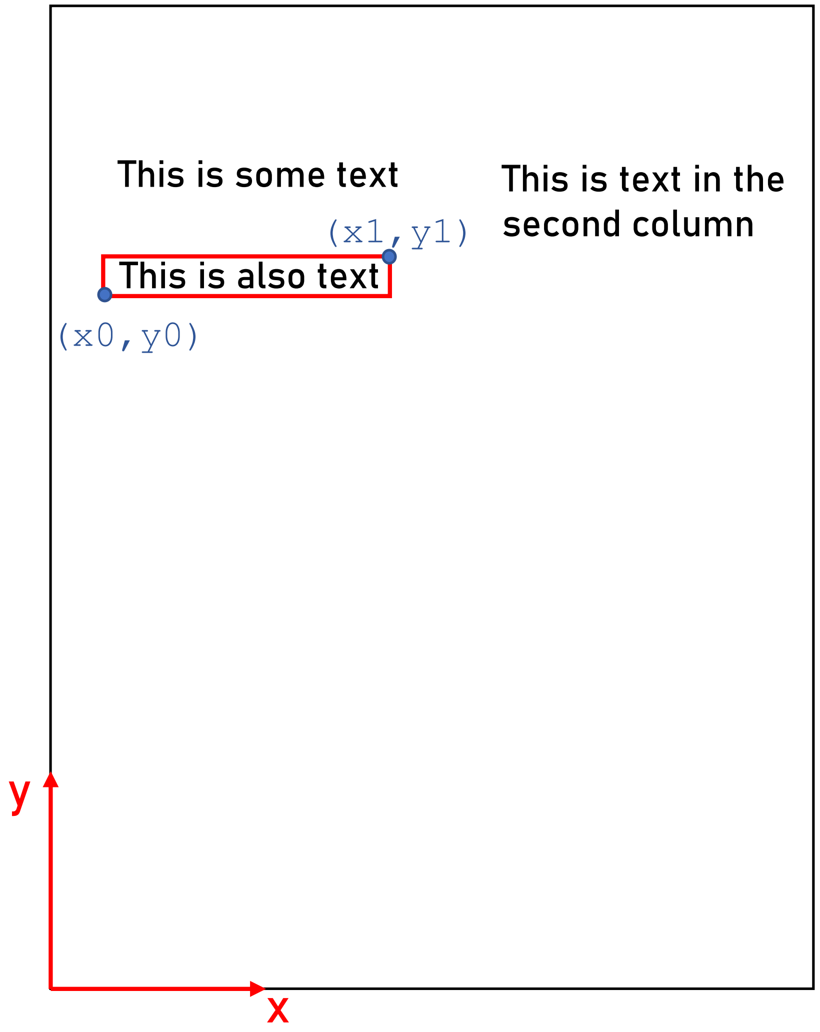
# Appendix: How it works – pdfminer



(page_nr, x0, y0, x1, y1, string)

Source: https://pdfminersix.readthedocs.io/en/latest/topic/converting_pdf_to_text.html

# Appendix: How it works – pdfminer



This is some text

This is also text

This is text in the second column

**What I extract:**

```
This is some text

This is also text

This is text in the
second column
```