

News Agency Classification

Annotation Guidelines

April 2023

Lea Marxen

In the following, you will find some guidelines and comments on how to annotate the corpus for the impresso News Agency Classification.

[1. News Agency Labels](#)

[1.1 News Agencies](#)

[1.2 The unk label](#)

[1.3 The pers.ind.articleauthor label](#)

[2. OCR noise](#)

[2.1 Noisy article](#)

[2.2 Noisy News Agency Mention](#)

[3. Token boundaries](#)

[3.1 Abbreviations](#)

[3.2 Compounds / Proper Names](#)

[4. Remaining Annotation Issues](#)

1. News Agency Labels

The News Agency Labels can be found in the layer "Impresso News Agencies".

1.1 News Agencies

Only annotate News Agencies when they provide the information contained in the articles. News Agency mentions which occur because the content is *about* the News Agency but not *from* them should not be annotated, e.g. if the article treats a change of personnel within a News Agency or an acquisition of a News Agency by another (and this information does not come from the News Agency itself), see also Fig. 1.

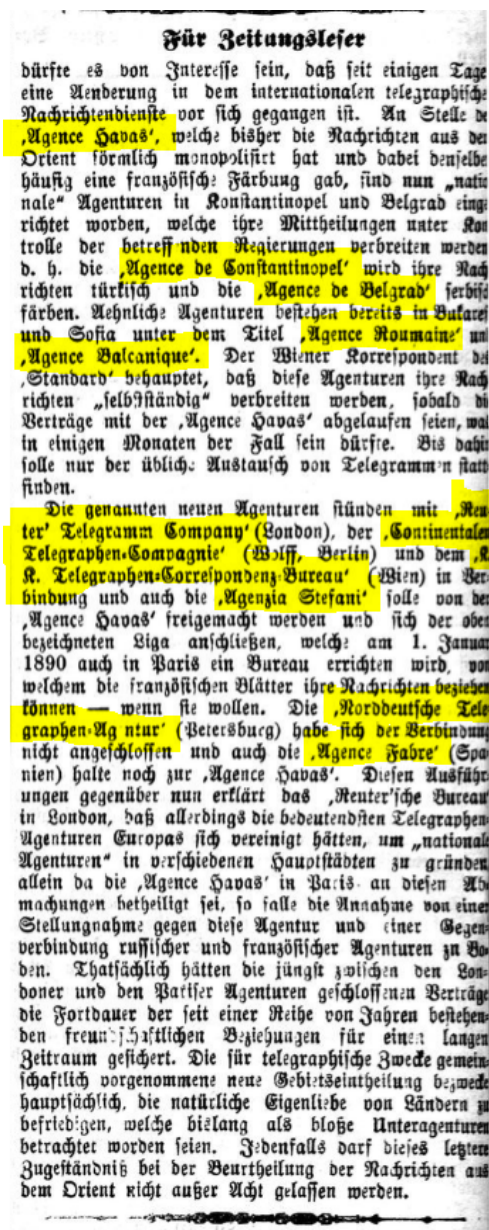


Fig. 1: Example for News Agency mentions which should not be annotated.

The following News Agency labels exist for annotation:

Abbr.	Other possible Reference In Newspapers	News Agency	Country	Creation	End
AFP	(AFP), (Afp.), (A. F. P.), Agence France Presse	Agence France-Presse	France	1944	
ANP		Algemeen Nederlands Persbureau	Netherlands	1934	
ANSA		Agenzia Nazionale Stampa Associata	Italy	1945	
AP	(AP)	Associated Press	USA	1848	
APA		Austria Press Agentur	Austria	1946	
ATS / SDA	Agence télégraphique suisse	Agence Telegraphique Suisse / Schweizerische Depeschagentur	Switzerland	1894	
BELGA	(Belga)	Agence télégraphique belge de Presse	Belgium	1920	
BTA	Agence Bulgare, BTA	Bulgarska Telegrafitscheka Agentzia / Agence Bulgare / Bulgarian Telegraph Agency	Bulgaria	1898	
CTK		Czechoslavenstva Tiskova Kancelar / Agence Ceteka	Czechoslovakia/Czech	1918	
DDP / DAPD	Ddp, dapd	Deutscher Depeschendienst / Deutscher Auslands-Depeschendienst	Germany	1971	2013
DNB	(D. N. B.)	Deutsches Nachrichtenbüro GmbH	Germany	1933	1945
DOMEI	(Domei.)	Domei Tsushin/Domei News Agency	Japan	1936	1945
DPA	(DPA)	Deutsche Presse Agentur	Germany (GFR)	1949	
Europapress	(Europapreß.), (Europapr.)	Europapress	Germany		
Extel	Agence Extel, (Extel.)	Exchange Telegraph Co. Ltd.	United Kingdom	1872	1993
HAVAS	(Havas.), Havas	Havas	France	1835	
Interfax		Interfax News Agency	Russia	1989	
PAP		Polska Agencja Prasowa	Poland	1944	
REUTERS	(Reuter), (Reuter.), Reutermeldung, Reuter'schen Bureau, Reuter-Telegramm	Reuters	United Kingdom	1851	
SPK / SMP		Schweizerische Politische Korrespondenz / Schweizer Mittelpresse (until 1947)	Switzerland	1917	1993

Stefani	Agenzia Stefani, (Stefani.)	Agenzia Stefani	Italy	1853	1945
TANJUG		Telegrafaska Agencija nova Jugoslavija	Yugoslavia/Serbia	1943	2018
TASS	ITAR-TASS, Taß, Telegraphen-Agentur der Sowietunion, Telegraphen-Agentur der U.S.S.R.	Telegrafnoie Agenstvo sovietskavo Soyusa	Russia	1925	
TT		Tidningarnas Telegrambyra	Sweden	1921	
Telunion	TU, (Telunion.)	Telegraphen-Union	Germany	1913	1934
UPI		United Press International	USA	1958	1990
Wolff	(Wolff.), Wolffbüro, Wolffsbüro, Kontinental-Telegraphie ncompagnie	Wolffs Telegraphisches Bureau	Germany	1849	1934

1.2 The *unk* label

The *unk* label should be used for News Agencies whose name is not contained in the predefined labels. If this is the case, please report it as a problem when finishing the annotation of the document and type “unk” in the message field:



Fig. 2: Press the indicated button to finish the annotation of a document.

Finish Document

This action will mark the document as **Finished**.

You can no longer make changes to the document after this step.

Only a project manager or curator can put the document back into editing mode.

If you indicate that the document had a problem, it will be removed from your list of documents and you can no longer access it.

Have you been able to successfully complete the annotations in this document?

☐ Success: annotation completed
☒ Problem: unable to complete the annotation

You can leave a message to the curator below.

unk

✓ Confirm

✗ Cancel

Fig. 3: Reporting a problem with message “unk” when finishing the document annotation

1.3 The *pers.ind.articleauthor* label

At some time, newspapers began crediting the authors of the articles. This can take the same form as crediting a News Agency, as can be seen in Figure 4. If you see this, use the *pers.ind.articleauthor* label to annotate the author of the article.



Fig. 4: Articles with News Agency credentials (yellow) and author credentials (red)

If you are unsure whether a reference belongs to a News Agency or the article other, label the token with *unk* and follow the procedure from [1.2](#).

2. OCR noise

There exist two different options to deal with OCR noise, one on document level to discard the article and one on token level to correct the spelling of a News Agency mention.

2.1 Noisy article

If a document is too noisy to annotate, you can indicate this by setting the label *non_usable* to “Yes” in the Document Metadata. For this, go to “Document Metadata” in the menu bar on the left, as indicated in Figure 5.

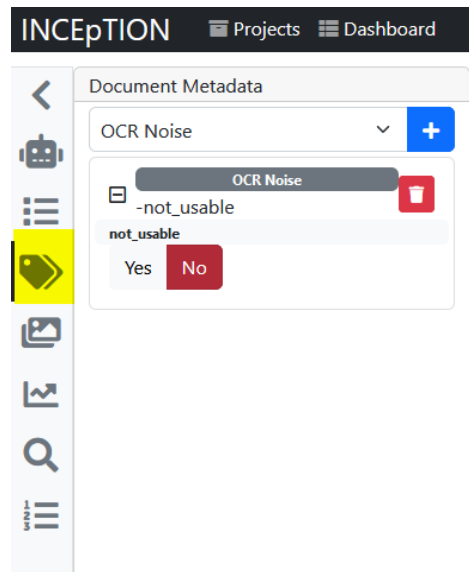


Fig. 5: Location of Document Metadata

2.2 Noisy News Agency Mention

If the OCR transcription of a News Agency mention is incorrect, please correct the transcription (and set the *noisy_ocr* tag to “Yes”).

Concerning the token boundaries, we will follow the specifications from the “Named Entity Annotation Guidelines” (HIPE):

Special cases with noisy OCR:

When it is difficult to establish the boundary of a mention because of noisy OCR:

- look at the image
- include, in the annotation, the garbage characters which you think should have been recognized and should be part of the mention
- mark the mention with the flag “noisy-entity” and add your OCR hypothesis correction.

ex: in the string Trève * (which stands for Trèves), the full string Trève * should be annotated, not only Trève.

3. Token boundaries

This section clarifies which part of a News Agency mention should be marked for annotation.

3.1 Abbreviations

Specification when to include “.” in annotation:

- **Include:** abbreviation of a name (e.g. “ag.”, “Ag. Télégr. Suisse”, “D.N.B.”)
- **Do not include:** at the end of an agency name (e.g. Havas. -> Havas) or the acronym without points in between (e.g. AFP. -> AFP)

3.2 Compounds / Proper Names

Do not include the words “agence” or “Agentur” in the annotation (e.g. agence **[Extel]**), except if they belong to a proper name (e.g. Agence France Presse, Agence Télégraphique Suisse).

Regarding German compounds, only include the name of the agency, e.g. **[Reuter]**meldung, **[Reuter]**-Telegramm.

Also note the following instructions from the “Named Entity Annotation Guidelines” (HIPE):

Special case with German compounds: Apply the cross-lingual or decomposition test, i.e. translate the compound to French and in the German compound annotate only what should be annotated in French.

The connecting “s” in German compounds is not annotated:

Völkerbundsmitgliedern

=> only Völkerbund is annotated

< org .adm> Völkerbund </ org .adm> smitgliedern

4. Remaining Annotation Issues

If it remains unclear how to annotate a certain mention, note the issue in the following file:

[Google Document for Annotation issues](#)

Resolving issues jointly and thus creating further examples of “best practice” can also ensure that the annotations will be as consistent as possible.