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.

Für Zeitungelefer burfte es bon Intereffe fein, baß feit einigen Tage eine Menberung in bem internationalen telegraphifd Radrichtenbienfte vor fich gegangen ift. An Stelle b "Agence Savas", welche bisher bie Rachrichten aus be Orient formlich monopolifirt hat und babei benfelbe häufig eine frangöfifch: Farbung gab, find nun "nati nale" Agenturen in Konftantinopel und Belgrad eing richtet worden, welche ihre Mittheilungen unter Kontrolle der betreffenden Regierungen verbreiten werden d. h. die Agence de Constantinopel' wird ihre Rachtichen türkisch und die Agence de Belgrad' serbist färben. Aehnliche Agenturen bestehen dereits in Bukare und Sosia unter dem Titel Agence Roumaine' und Agence Balcanique'. Der Wiener Korrespondent die Standard' behauptet, daß diese Agenturen ihre Rachtichen "selbsständig" verbreiten werden, sobald di Berträge mit der Agence Havas' abgelaufen seien, wal in einigen Monaten der Fall sein dürste. Bis dahit solle nur der üblich. Austausch von Telearammen statt richtet worben, welche ihre Mittheilungen unter Ron folle nur ber üblich. Mustaufch von Telegrammen ftatt finben. Die genannten neuen Agenturen ftunben mit ,Ren ter' Telegramm Company' (Bondon), ber Continentales Telegraphen Compagnie' (Bolff, Berlin) und bem ,L R. Telegraphen:Correspondenz. Bureau' (Bien) in Ber bindung und auch die "Agenzia Stefani' solle von ber Agence Havas' freigemacht werden und fich der ober bezeichneten Liga anschließen, welche am 1. Januar 1890 auch in Baris ein Bureau errichten wirb, von welchem bie frangofischen Blatter ihre Rachrichten besieben tonnen wenn fie wollen. Die "Rordbeutsche Tele graphen Ag ntur" (Betersbucg) habe sich der Berbindung nicht angeschlossen und auch die "Agence Fabre" (Spa nien) halte noch jur "Agence Savas". Diefen Ausfuhr ungen gegenüber nun ertlart bas "Renter'iche Burean ungen gegenwer nun ertiatt das "keinerige Gureau in Bondon, baß allerdings die bedeutendften Telegraphen Agenturen Europas fich vereinigt hätten, um "nationale Agenturen" in verschiedenen Hauptstädten zu geinden allein da die "Agence Havas" in Bacis an diesen Abmachungen hetheiligt sei, so salle die Annahme von einer Stellungnahme gegen diese Agentur und einer Gegen verbindung rufflicher und französischer Agenturen zu Bo ben. Thatfachlich hatten bie jungit swifchen ben Son-boner und ben Battfer Agenturen geichloffenen Bertrage bie Fortbauer ber fett einer Reibe von Jahren bestehen freun' faftlichen Begiehungen für einen langen Beitraum gefichert. Die für telegraphifche 3mede gemein chaftlich porgenommene neue Gebietseintheilung beimed hauptfachlich, bie natürliche Gigenliebe von ganbern m befriedigen, welche bielang als bloge Unteragenturen betrachter worden feien. Jebenfalls barf biefes legten Bugeftandnig bei ber Beurtheilung ber Radrichten aus bem Drient nicht außer Acht gelaffen merben.

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	France	1944	
ANP	Agence i fance i fesse	Algemeen Nederlands Persbureau	Netherlands	1934	
ANSA		Agenzia Nationale Stampa Associata	Italy	1945	
7114071		Agenzia Nationale Stampa Associata	ricity	1040	
AP	(AP)	Associated Press	USA	1848	
APA		Austria Press Agentur	Austria	1946	
ATS / SDA		Agence Telegraphique Suisse / Schweizerische Depeschenagentur	Switzerland	1894	
BELGA	(Belga)	Agence télégraphique belge de Presse	Belgium	1920	
вта	Agence Bulgare, BTA	Bulgarska Telegrafitscheka Agentzia / Agence Bulgare / Bulgarian Telegraph Agency	Bulgaria	1898	
СТК		Czechoslavenska 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	
Europapre ss	(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 Büreau, Reuter-Telegramm		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		Telegrafska 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	Russia	1925	
TT		Tidningarnas Telegrambyra	Sweden	1921	
Telunion	TU, (Telunion.)	Telegraphen-Union	Germany	1913	1934
UPI		United Press International	USA	1958	1990
	(Wolff.), Wolffbüro, Wolffsbüro, Kontinental-Telegraphe			40.60	100 1
Wolff	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.

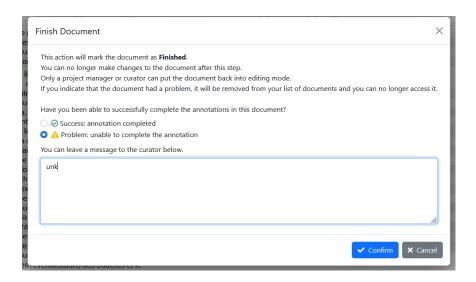


Fig. 3: Reporting a problem with message "unk" when finishing the document annotation

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.

ÉTATS-UNIS Grâce refusée à Troy Davis

La justice américaine a refusé mardi d'accorder sa grâce à Troy Davis, un Noir condamné à mort en 1991 pour le meurtre d'un policier blanc et devenu un symbole de la lutte contre la peine de mort. Cette décision tombe à la veille de son exécution prévue dans l'Etat de Géorgie. «Le comité a refusé sa clémence», a indiqué dans un communiqué le comité des grâces de Géorgie. La réunion de ce comité à Atlanta, la capitale de Géorgie, était considérée comme la dernière chance pour le condamné de voir sa peine de mort commuée en prison à vie, le gouverneur de l'Etat ne disposant pas du droit de grâce. L'exécution de Troy Davis par injection mortelle est programmée jeudi à 1h en Suisse à la prison de Jackson, malgré des doutes sur sa

FRANCE Vente de manuscrits de Gainsbourg

Les manuscrits des chansons «Sorry Angel», «Love on the beat», «No Comment», «Hmm Hmm Hmm» et de «You're under arresto seront mis en vente par Sotheby's Paris le 9 novembre prochain. Des notes diverses, des tapuscrits ainsi que des photos inédites de Serge Gainsbourg (1928-1991) réalisées pour un magazine anglais compléteront cette vente. Le brouillon manuscrit de «Love on the beat» (1984) est peut-être le plus fascinant. Selon les organisateurs de la vente, ces deux pages comportent de très nombreuses variantes et corrections faisant apparaître en filigrane les influences baudelairiennes présentes dans toute l'œuvre de l'artiste (estimation: entre 12 000 et 18 000 euros). O PHV

NIDWALD Un soldat blessé lors d'un exercice

La justice militaire enquête depuis un mois sur un incendie qui a fait un blessé lors d'un exercice à Oberdorf (NW), près de Stans Souffrant de blessures légères à moyennes, ce soldat en cours de répétition avait recu fordre d'allumer un feu destiné à être éteint par des soldats au centre de formation de la Swisscoy. Les faits se sont produits le 23 août demier a indiqué hier la justice militaire. Le blessé n'était pas membre de la troupe qui devait être envoyée au Kosovo. Motocydiste, il faisait partie d'un détachement responsable de l'exploitation des installations du centre de formation Swissint. Les circonstances de l'accident font fobjet d'une enquête L'investigation en est à son stade initial. Des spécialistes de l'institut médico-légal de la ville de Zurich se sont joints aux enquêteurs. O ATS

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 $\underline{1.2}$.

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.

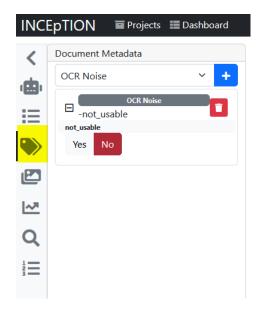


Fig. 5: Location of Document Metadata

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.