



Bibliothèque nationale
du Luxembourg

Fine grained language identification in multilingual corpus with OCR errors

**CENL Network Group “AI in
Libraries” Webinars 2023**

30.5.2023 14:00 CET

Yves Maurer
Deputy Head of IT
National Library of Luxembourg



National Library of Luxembourg



Founded in 1789
New building: 2019

Heritage, public and
university library

40. 000 users
More than 1.8M printed
items
75.000 ejournals, 900.000
ebooks
Head of network of 89
libraries
1.2 Million digitised pages



eluxemburgensia.lu

2003-2006

- Image only
- Newspapers, postcards, books, manuscripts

2006+

- Luxembourg newspapers
- Image & rich metadata (METS/ALTO)

2015

- 15 medieval manuscripts with metadata
- 441 books with METS/ALTO

2016+

- Newspapers, books, postcards etc.

Layout segmentation

- Issue > Section > Article > Paragraphs
- Search works on articles

SECTION

Intérieur.

A TRAVERS L'EXPOSITION D'ETTELBRUCK.

ARTICLE

Dimanche passé, Ettelbruck a commencé ses fêtes. L'entrée dans la ville où l'on était alors et givré, curieux à la source de musique de Diekirch, notre petite ville présentait un aspect assez animé. Mme par un sentiment d'excuse convenable, cette noble société de Diekirch tenait à participer à l'ouverture de nos fêtes; merci pour son attention!

Donc le grand jour, le jour du festival approche; on met la dernière main aux préparatifs; le travail avance lentement; le marché respire, sûr qu'il est de pouvoir finir à temps. En attendant, voyons un peu ce qui est fait.

L'exposition, me dira-t-on? Oui, l'exposition. L'idée a pris corps; notre exposition existe; nombre de visiteurs en ont et auront l'occasion de l'admirer. Mais préférons avec ordre!

Le tour du rez-de-chaussée est fait; montons, pour faire pareille promenade au premier. Entrons dans cette première salle à droite. Vous la voyez, les ébenistes et les menuisiers en font les honneurs. Admirez donc ces jolis meubles de salon, ces armoires, ces commodes, ces tables, ces fauteuils, ces chaises, ces bois de lit, etc., etc. Ne dédaignez ni ces mignons modèles d'escaliers à échelle réduite, ni ces dessins qui tapissent les murs. Et pour peu que vous soyiez amateur de musique, examinez et essayez ces jolis pianos, ces beaux instruments en cuivre. Tout cela est parfait et la salle est belle!

Dès deux côtés de la belle place devant la maison d'école, s'élève le kiosque dont mention à plusieurs reprises. Quoiqu'encore entouré d'échafaudages, il fait bonne mine par son élégance, sa forme svelte et élancée. C'est l'œuvre de l'architecte M. Knepper de Diekirch.

Des deux côtés de l'école et comme faisant fond à ce bâtiment, on a élevé des constructions en bois, qui ferment la place du côté de l'Alzette. Le tout forme un ensemble assez coquet; ce sont les locaux de l'exposition.

Pronons à gauche, c'est là l'entrée. Par un tourbillon nous arrivons dans une cour où plié un petit jardin anglais avec jet d'eau au milieu, terres en fleurs et frais gazon tout autour et une belle collection de mosaïques le long du mur de la maison d'école.

Le portail devant nous nous entra à la

rez examiner tout en entrant, vous ne voyez tout autour de la salle que des dessins de différents genres, albums, livres, etc. Le fond de la salle est occupé par une belle collection de tableaux et de portraits dus au pinceau d'un artiste dont je n'ai pas besoin de citer le nom; le nombre de ses collègues est restreint dans notre pays. Salon très-intéressant! Salon qui mérite d'être vu! —

N'y restons cependant pas; passons plutôt par cette porte du fond et nous arriverons de plein pied dans les produits de la tannerie, de bon et de beau; vous seriez difficile pour ne pas trouver votre fait. Et des chaussures? Amenez, Madame; elle vous dira s'il y en a beaucoup, de fortes et de belles! Mais n'oubliez pas la sellerie à gauche; peut-être vous faudra-t-il ces jolis harnais pour atteler votre pony à une des élégantes voitures de tantôt. — Salon bien remarquable que celui des cuirs et des chaussures!

Le tour du rez-de-chaussée est fait; montons, pour faire pareille promenade au premier. Entrons dans cette première salle à droite. Vous la voyez, les ébenistes et les menuisiers en font les honneurs. Admirez donc ces jolis meubles de salon, ces armoires, ces commodes, ces tables, ces fauteuils, ces chaises, ces bois de lit, etc., etc. Ne dédaignez ni ces mignons modèles d'escaliers à échelle réduite, ni ces dessins qui tapissent les murs. Et pour peu que vous soyiez amateur de musique, examinez et essayez ces jolis pianos, ces beaux instruments en cuivre. Tout cela est parfait et la salle est belle!

La porte à gauche conduit dans une pièce qui ne compte pas de nom général. Les objets y exposés sont de nature trop diverse pour être rangés dans une catégorie. Ainsi, à droite vous voyez des machines à coudre, devant vous un étalage de confection, à gauche quantité de casquettes, de bonnets, etc., au fond une belle collection de draps. Au million

vés à la sortie, au terme de notre promenade. Somme toute, l'exposition est belle, réussie, même réussie qu'en osait l'espérer.

Elle sera visitée, je n'en doute pas. Dans cette conviction, je m'en vais prendre une chope, mais sans bruit; la société Ettelbruck est morose; elle prépare d'après ce qu'AUTHOR

X Y.

Nous apprenons avec plaisir que le bureau exécutif de l'Exposition nationale, qui aura lieu prochainement à Bruxelles, à l'occasion des fêtes du cinquantenaire de l'indépendance, a appelé M. Charles Buffet, secrétaire de la Société royale agricole, à Wiltz, pour faire partie du jury pour le concours de reprojet de loi.

ARTICLE

PROJET DE LOI

concernant les traitements et pensions des secrétaires communaux.

Art. 1^{er}. — Le traitement des secrétaires communaux est fixé par le conseil communal sous l'approbation du Gouvernement, au minimum dans la proportion de 30 à 35 centimes par tête de population de la commune.

Il pourra être accordé, sur les fonds de l'Etat, aux communes, en regard à leur situation financière, des subсидies pour couvrir une partie de l'augmentation de traitements pouvant résulter de l'application de la disposition qui précéde. Ces subсидies ne pourront dépasser le chiffre annuel de 4000 francs.

Art. 2. — Le Gouvernement pourra allouer aux secrétaires communaux, sur les fonds de l'Etat, après cinq ans de bons services remplis conséutivement dans la même commune, un supplément de traitement qui est fixé comme suit:

40 francs dans les communes dont la popula-

Ne compte pendant lequel remplies par su

Art. 5. — I droit à la pensi 1^{er} s'il abando avant d'en av

sionné;

2^{er} s'il est con ou à une peine an, on à l'intér

fonction, emplo

pensionné. Il pa

tant les disposi

16 janvier 1865

Art. 6. — Chaque année d

traitemen moye

Si le service

la moyenne ser

fective.

Le temps de

et par mois; les

Art. 7. — I

rieure à 150 fr

du traitemen

Art. 8. — D

de l'art. 3, la p

Cette majorati

50 pCt., si l'in

cident, des pre

ment extraordi

Dans aucun

être au-dessous

trois quarts du

base.

Art. 9. — S

fonctions success

plusieurs comm

d'autant de part

et chaque part

double base du

de service pass

ment à l'art. 6.

Le maximum

B

Multilingual newspapers

**MAISON
BERNHARD-SCHROEDER
SUCCESEUR
VINS ET LIQUEURS**
Obercorn,
fermée du 22 juin au 2 juillet
7294

**Achtong !
Ve'hbesetzer a Baueren.**
No 5jähregem Exil erem dohém, délen ech de Ve'hbesetzer mat, dass ech erem ewe' fre'er mat NOTZ- an ZUCHTVE'H den Handel weiterbedreiven.
Zo' gleicher Zeit délen ech Iech mat, dass ech vum Stat. am Hollerecher Schluëchthaus als Ve'hkommissionär ernannt sin. Dir könnt Ert fett Ve'h bei mir umellen. 4469
An der Hoffnung, daß der mir Ert Vertraue schenkt ewe' emmer, gre'ßen ech Iech mat aller Hochachtong

Den Israëls Décken
BO'NEWEG,
Letzeburgerstroß N°. 65.
Tel.: Schluëchthaus 58-32.

Hère-Coiffeur-Salon
opmachen.
Em gene'gten Zo'sproch biéd den Edmond THULL.

FOURRURES
Maison de Gros - cherche clients pour le Luxembourg.
Ecrire: Armand GRAULS,
Bruxelles, rue Grétry 26.
4545

OUVERTURE
à ESCH-Alzette

School of Languages
Helen WIES-YULL
(engl. Teacher)
ENGLISH - FRENCH
FRENCH and GERMAN
for Allied Troops
ESCH, 66, Av. de la Gare.
Inscriptions pour les cours:
les mercredis de 2 à 7 heures
les samedis de 3 à 5 heures
4225

RAJEUNISSEZ !
Redevenez souple et alerte comme à 20 ans, en éliminant l'excès d'acide urique accumulé dans votre orga-

traueren d'UNIO'N an de POMPIERSCORPS v. He'en.
4557

Marcel NEISELER, Zenn-techniker, gefal fir seng le'w Hémecht bei Kiel, den 3. Mé 1945, am ble'enden Alter vun 18½ Jor. Leichendengscht e Mettwoch, de 27. Juni, em 10 Auer, zo' Scheffleng. 7327
Famill Neiseler-Ludovicy.

Sisy SCHUH, gestuerwen durch Fliegerugreiff am K. H. D. Pforzheim, den 23. Februar 1945, am Alter vun 20 Jor. Feierleche Leichendengscht e Méndeg, de 25. Juni, em 10.10 Auer zo' Kéi, 7301 Familjen Schuh-Herzig.

Jengy EDERT, Man vum Madeleine MICHAELY, gestuerwen fir seng Hémecht am KZ. Dachau, den 8. Abröll 1945, am Alter vu 55 Jor. Feierleche Leichendengscht e Mettwoch, de 27. Juni, em 10 Auer zo' Pé'teng. 4578

Remerciements
Kun HURST, anc. officier du génie aux Indes Néerlandaises. La messe de six semaines sera célébrée le lundi, 25 juin, à 10 heures, en l'église St. Joseph, Esch-Alzette. Merci spécial pour les belles fleurs et les saines messes. 4521



Why identify the language?

- Interesting by itself for multilingual country
- Full text search
- Better OCR by also using language specific dictionaries
- AI Projects since 2020
 - Topic modelling
 - Transformers
 - ...

B Starting out

- We knew that languages are:
 - German (roughly 60%)
 - French (~ 40%)
 - Luxembourgish (~ 1%)
 - English (~ 0.1%)
 - Unknown long tail of other languages
- Need for method which supports Luxembourgish

Luxembourgish

- Luxembourgish case is hard since corpus is 1841 – 2010 and spelling was only standardized in 1975.

- 1846 Wann ech an 't Grós-Gaas 'lo erem kém
- 1869 dé well Schwein hu glât këng
- 1947 De'w an onst Hierz
- 2007 deen am *Cactus* kengTut méi kritt

B

Challenges

- OCR errors
- Multilingual content inside a single article
- Articles without language (e.g. train time tables, sports results, ...)
- Very uneven distribution of languages
- How to identify long tail of minority languages
- How to limit false positives
- Luxembourgish is often not supported, e.g. [langdetect](#), [lingua-py](#)
- Limited time-frame : 3 months (project team of 2)

Features

- 3 base algorithms
 - CLD3 (google)
 - Langid (University of Melbourne)
 - Fasttext (lid218e from
<https://github.com/facebookresearch/fairseq/tree/nllb#lid-model>)
- Stop words for Luxembourgish
- Dictionaries for each language
- Garbage Tokens
- Uppercase letters
- Length of text
- Newspaper identifier

Feature evaluation process

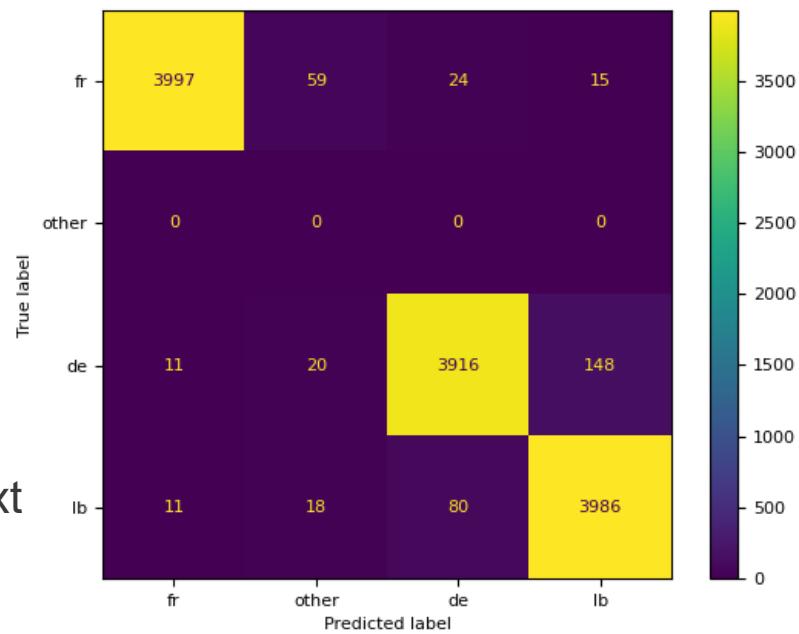
- Extract full-texts
- Compute all features
- Load into elasticsearch
- Evaluate using Kibana
- Extract lists to csv



Compare cld3, fasttext & langid

- First try with ground truth set (4095 * 3)
 - French, German, Luxembourgish
- Best: 97.4% Accuracy
 - langid+LB stopwords
- Disadvantages
 - Real data is skewed
 - Error rate dwarfs true positives
 - No GT for all languages

Confusion matrix for fastText



Combine : Majority vote

- If fastText == CLD3 -> fastText, otherwise langid
- Test on 100 000 articles. Investigate 300 cases where langid gets outvoted

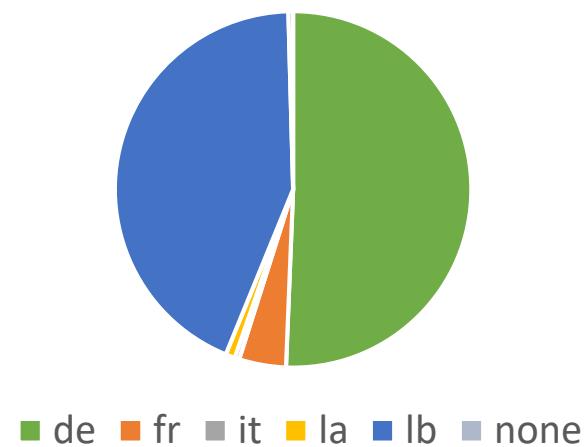
	de	fr	lb	da	en	nl	zh	jv	br	eo	eu	fi	la	et	oc	ga	sv	pl	sk	no	it	cy	ht
de	30	17	49	52	49	9	1	2	2	0	1	2	3	1	1	0	5	2	1	1	0	1	0
fr	1	14	15	2	2	0	0	0	0	1	0	0	0	0	2	2	0	0	0	1	1	0	0
lb	8	6	3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Voting	de	fr	lb	da	en	nl	af	capt	vi	hu	it	cs	
de	181	7	22	3	2	6	6	0	2	0	1	0	1
fr	9	24	8	1	1	0	0	2	0	2	0	1	0
lb	2	0	16	0	0	2	0	0	0	0	0	0	0

Should we use stopwords?

- Random sample of 100 000 articles
 - real distribution of languages
- Check manually all 235 cases when stopwords overrule langid
 - Aim: avoid false positives (LB is only around 1% of corpus)

	#correct	Accuracy
stopwords for Luxembourgish	102	43.22
Langid	109	46.19
CLD3	173	73.31
fastText	183	77.54





Identify “minority” languages

- Examine cases langid=cld3=fasttext and not de, fr, lb or en
- 3317 cases from 8 million reviewed in Excel

algo-new	cld3	fasttext	langid	id	LINK	correct other	textStart		
af	af	af	af	http	https://persist.l		TEKKA, die ideale KrankenverH		
it	it	it	it	http	https://persist.l	it	I celebri prodotti itniani alia po		
nl	nl	nl	nl	http	https://persist.l	nl	Vcor ons kantoor in Luxembur		
nl	nl	nl	nl	http	https://persist.l		Generalstaatsanwalt Louis Sab		
af	af	af	af	http	https://persist.l		Josiane Graas (T.C. Arquebusie		
es	es	es	es	http	https://persist.l	es	Embajada de Espana Informac		
ro	ro	ro	ro	http	https://persist.l		Ã®louA v-oi cadeaux utilei		
it	it	it	it	http	https://persist.l		Sur notre photo le Trio dâ€™Ai		
es	es	es	es	http	https://persist.l	es	EMBAJADA DE ESPANA Rectific		
ro	ro	ro	ro	http	https://persist.l		Octane de Marcus Adams		
es	es	es	es	http	https://persist.l	es	Embajada de Espana Informac		



cld3=fasttext=langid agreement on minority languages

		by agreement	correct	%
Italian	it	427	285	66.74
Portuguese	pt	77	50	64.94
Dutch	nl	263	38	14.45
Polish	pl	98	43	43.88
Spanish	es	47	10	21.28
Danish	da	37	2	5.41
Hungarian	hu	30	5	16.67
Esperanto	eo	9	6	66.67
Bosnian	bs	1	1	100
Croatian	hr	4	3	75

Computing other features

- Fix list of allowed languages and use dictionaries to compute $\text{len(tokens in dict)} / \text{len(tokens)}$
- Detect bad OCR using heuristics
- Count uppercase letters
 - langid tends to favor to English for Uppercase strings
- Length of text
 - Short texts are ambiguous

B

Final algorithm

- General case:

If chars < 12: unknown

If garbage > 50%: unknown

If cld3=fasttext: x=cld3 else x=langid

- Per language, specific rules, e.g.:

If chars < 40 or dict < 50%: unknown

- For really small languages, have a fixed list

Minority languages – filtering data

- Use articles found previously to visualize their properties in Kibana
- Determine filter that allows all GT articles through
- E.g. Polish:

Ground Truth: 85

Algorithm: 1635

Filter: Dict > 30% and Uppercase < 70% and chars > 60 and Garbage < 20%

Filtered: 242

- Validate the 242 articles, found 83 more

Results of filtering

Language	algo	Filter	Ground Truth	Additional	Total real	Percentage from algo	Percentage from Filter
Portuguese	4309	298	111	100	211	4.9	70.81
Polish	1635	242	85	83	168	10.28	69.42
Dutch	20720	769	52	27	79	0.38	10.27
Spanish	4825	226	20	12	32	0.66	14.16
Esperanto	753	25	11	4	15	1.99	60
Hungarian	2676	45	8	6	14	0.52	31.11
Croatian	993	9	3	5	8	0.81	88.89
Ido	0	0	2	2	4	0	0
Danish	6530	32	2	0	2	0.03	6.25
Bosnian	781	9	2	0	2	0.26	22.22
Gaelic	879	21	2	0	2	0.23	9.52
Slovenian	1848	46	1	0	1	0.05	2.17

Going into production

- Integrated into METS – MODS

```
<dmdSec ID="MODSMD_MAP4">
  <mdWrap MDTYPE="MODS" MIMETYPE="text/xml">
    <xmlData>
      <mods:mods>
        <mods:titleInfo ID="MODSMD_MAP4_TI1" xml:lang="de">
          <mods:title>Abb. 1. Auswahl der neun repräsentativen Rasterquadrate.</mods:title>
        </mods:titleInfo>
        <mods:language>
          <mods:languageTerm authority="rfc3066" type="code">de</mods:languageTerm>
        </mods:language>
      </mods:mods>
    </xmlData>
  </mdWrap>
</dmdSec>
  ...
  ...
  ...
```

- Determined language for Manuscripts and Posters (no OCR) from catalogue data
- Set book language to catalogue data



Online since 27.4.2023

Search Facet

- German (5261387)
- French (2459078)
- Luxemburgish (118123)
- English (10149)
- Latin (1959)
- Italian (1379)
- Portuguese (212)
- Polish (168)
- Dutch (78)
- Spanish (32)
- Esperanto (15)
- Hungarian (14)
- Croatian (8)
- Ido (4)
- Danish (2)
- Irish (2)
- Bosnian (2)
- Russian (2)
- Slovenian (1)



Sterbefälle

Article • Bürger- und Beamten-Zeitung • Wednesday, 24 December 1902



Publicité 5 Page 3

Advertisement • Bürger- und Beamten-Zeitung •
Wednesday, 24 December 1902



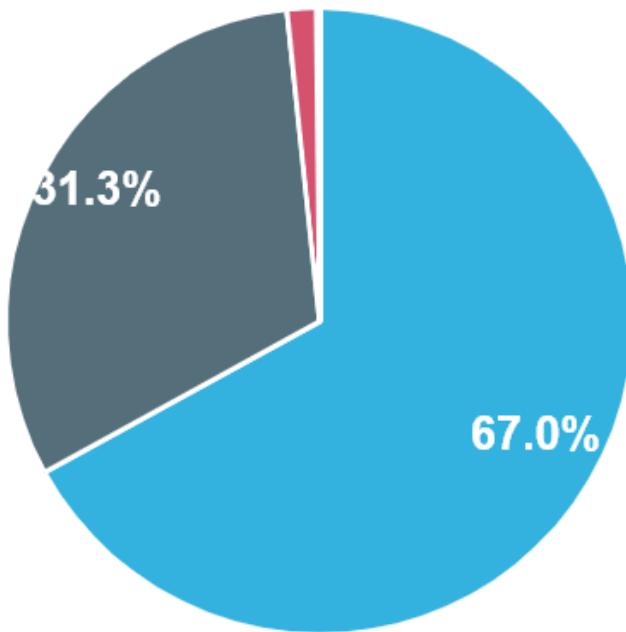
Heitere Ecke.

Article • Bürger- und Beamten-Zeitung • Wednesday, 24 December 1902



Languages in eluxemburgensia

LANGUAGE DISTRIBUTION

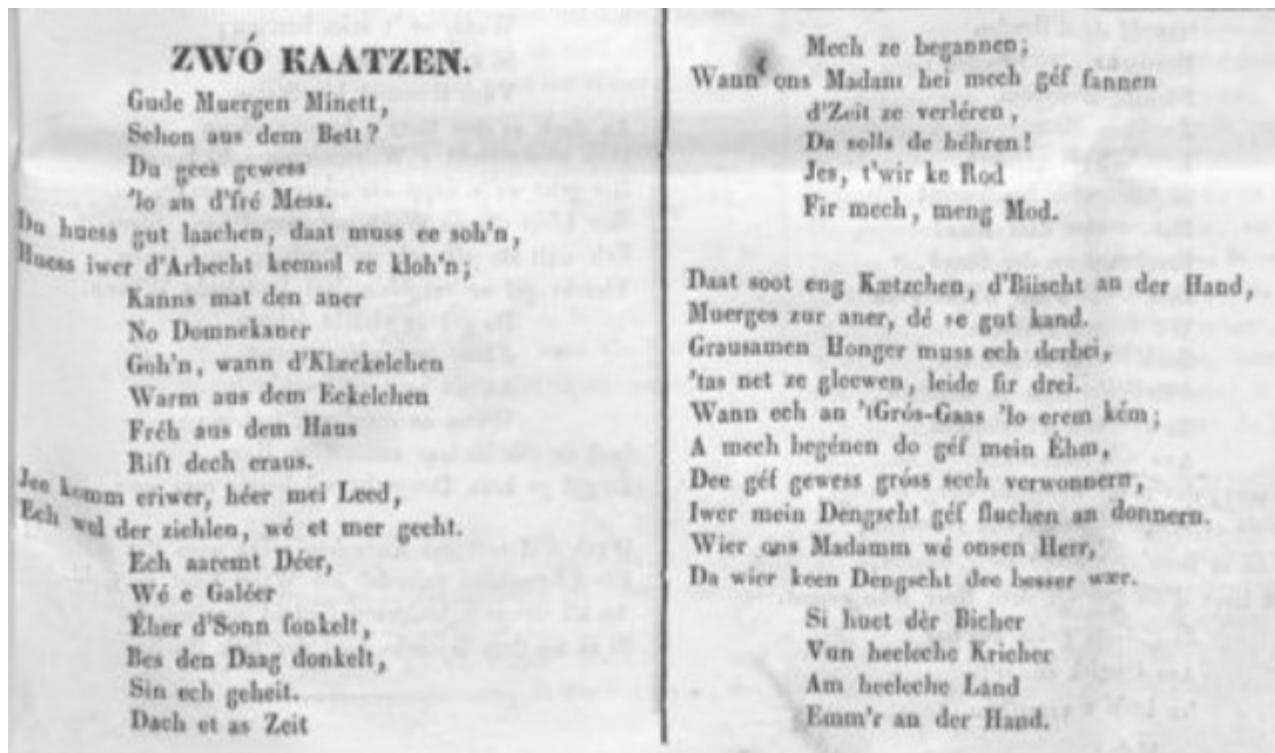


- German (5261387) ● French (2459078) ● Luxemburgish (118123) ● English (10149) ● Latin (1959)
- Italian (1379) ● Portuguese (212) ● Polish (168) ● Dutch (78) ● Spanish (32) ● Esperanto (15)
- Hungarian (14) ● Croatian (8) ● Ido (4) ● Russian (2) ● Bosnian (2) ● Irish (2) ● Danish (2)
- Slovenian (1) ● af (0)

Interesting finds

First Luxembourgish (1846) in the collection

<https://persist.lu/ark:70795/83sbwv/pages/1/articles/DTL35>



Finds : Immigration from Italy

1882: <https://persist.lu/ark:70795/89dsqk/pages/3/articles/DTL55>

Briefkasten.

Sign. Ha., Luss. Dica alla Sua sorella Ang. che oggi arriva alla bella età di anni 34. Adesso o mai !

1904: <https://persist.lu/ark:70795/g320c1/pages/4/articles/DTL49>

La pagina dei fratelli italiano.

Guerra —

I campi seminati di morti, i lugubri ossari eretti alla memoria delle innocenti vittime non valgono a distogliere i due potenti imperi Russo e Giapponese dal pigliare le armi. La diplomazia europea che assiste fredda e unita allo scannarsi di due popoli non può comprendere quanti dolori, quanti lutti, quante lagrime farà versare la pazzesca ambizione di due potenti. 2 miliardi 500 milioni non meno costerà

Quanta forza, quanta energia, quante vite umane! Ed il popolo che piange i suoi cari, il popolo che piange i suoi figli, mira e tace a tale scenario. Quando il torpore che lo avvince scomparirà? Quando in uno slancio sublime di fraterna solidarietà spezzerà le catene che lo tiene avvinto? Rispondiamo noi.

Quando le camorre, le mafie, i tristi connubii, le schifose alleanze saranno scomparse; quando l'istruzione, la scienza avrà fatto scomparire tutte le idiote superstizioni di amor patrio di bandiere di nazionalismo, allora la guerra sarà diventata impossibile. Povero Tolstoi! povero Zolas! quanto avete fatto per combattere quest' Idra! G. I.

Togliamo dal „Grido di Milano“.

Una sepolta Viva.

A San Pietro presso Cava dei Tirreni fu scoperta una povera donna sepolta viva da quindici anni in un sotterraneo. Essa fu sequestrata da due suoi fratelli — preti — per carpirne l'eredità paterna. Il fatto ci fa inorridire. Oltre le angosce d'una donna sotterrata viva per quindici lunghi anni, vi è l'infamia dei due predicatori di Cristo, o l'avidità dell'oro.

Non bastava il Medioevo tenebroso, la strage degli Albigesi, i misteri dell'Inquisizione ecc.; non bastavano i misteri dei conventi, le funzioni sacerdotali, le inversioni ripugnanti, i delitti feroci e raffinati, tutta l'aureola gloriosa del prete: oggi — proprio oggi — secoli di civiltà e progresso, due jene, che salmodiano laudi al Signore — bontà e carità — compiono le azioni più orrende che non sono spiegabili dalla passionalità dell'uomo o dall'ascetismo cieco, ma solamente da un istinto di lucro che li rende freddi per quindici anni, e non lascia loro un sussalto, né un ravvedimento.

La pietà, la carità, la rassegnazione che questi

R. È un giornalotto.

P. Un giornalotto!

R. Sì, il „Seme“.

P. Oh bella, non l'ho mai inteso nominare; e di che tratta?

R. Ecco mi spiego in due parole: Questo giornalotto sortì a Terni per iniziativa del compagno Paoloni, poi morì per . . . ora si è preso l'incarico il compagno editore Mongini come fece per altri giornali, perciò speriamo che non muoia più.

P. A quale scopo serve questo giornale?

R. Capita porta il nome con sé; serve per seminare nella mente digiana ed offuscata del proletariato le idee del Socialismo, per la emancipazione, ossia per la trasformazione della decrepita e barcollante società attuale e

P. Scuratamente quanto costa?

R. Un centesimo in Italia, e due all'Estero.

P. Costa una mia chionerla, vendimene uno.

R. O questo solo; è un numero di saggio che ha mandato l'editore . . .

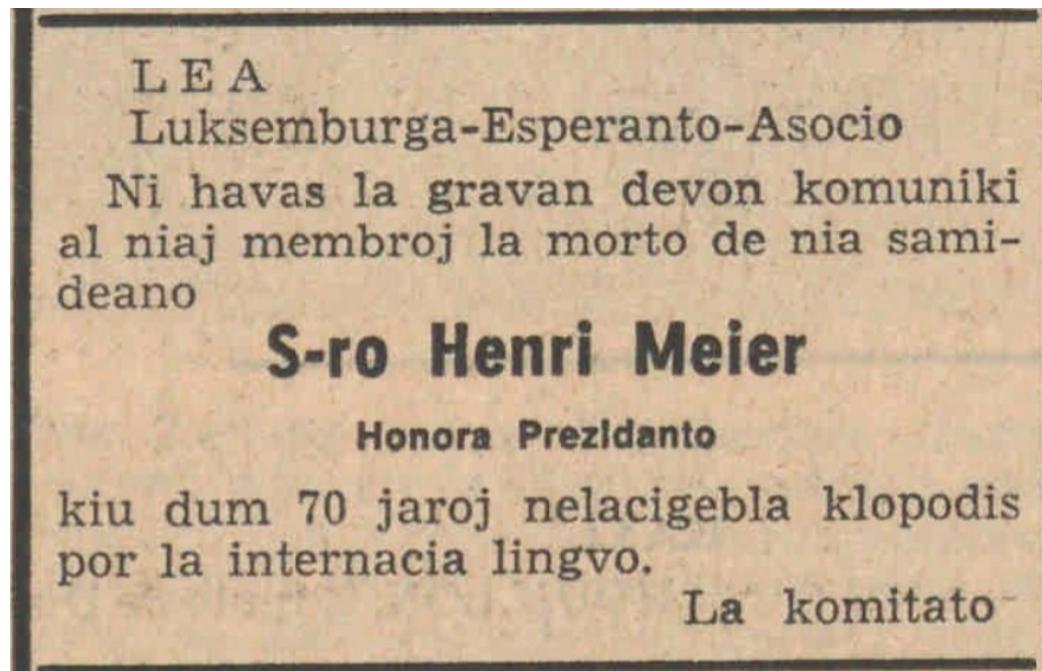
P. Per . . .



Finds: Esperanto

Death notice in Esperanto

1973 <https://persist.lu/ark:70795/hwpvbp3hq/pages/12/articles/DIVL2016>



B Finds: English textbook controversy

First article in English

1880: <https://persist.lu/ark:70795/qhbcqf/pages/2/articles/DTL56>

To Mr. F. Berger, formerly a Banker,
now a Deputy and Consul-General.
(Residences: America, Paris, Luxemburg.)

Sir,

Since you declare (and no one can possibly
be a better judge than yourself!) that my im-

obedient servant.

Dr WECKERING,
Professor.

NB. In order to prevent you from getting your
an-wer translated into English by some oth r person,
I beg you will reply within the space of twice
twenty four hours, which is but fair, as I always
answered your articles within half that time from
their appearance.



Finds: Hungarian

23.10 - 4.11.1956: Hungarian Uprising

22.12.1956: Christmas services for refugees

<https://persist.lu/ark:70795/3rz30pqx1/pages/5/articles/DIVL404>

Karácsonyi istentisztelet a magyar menekültek számára

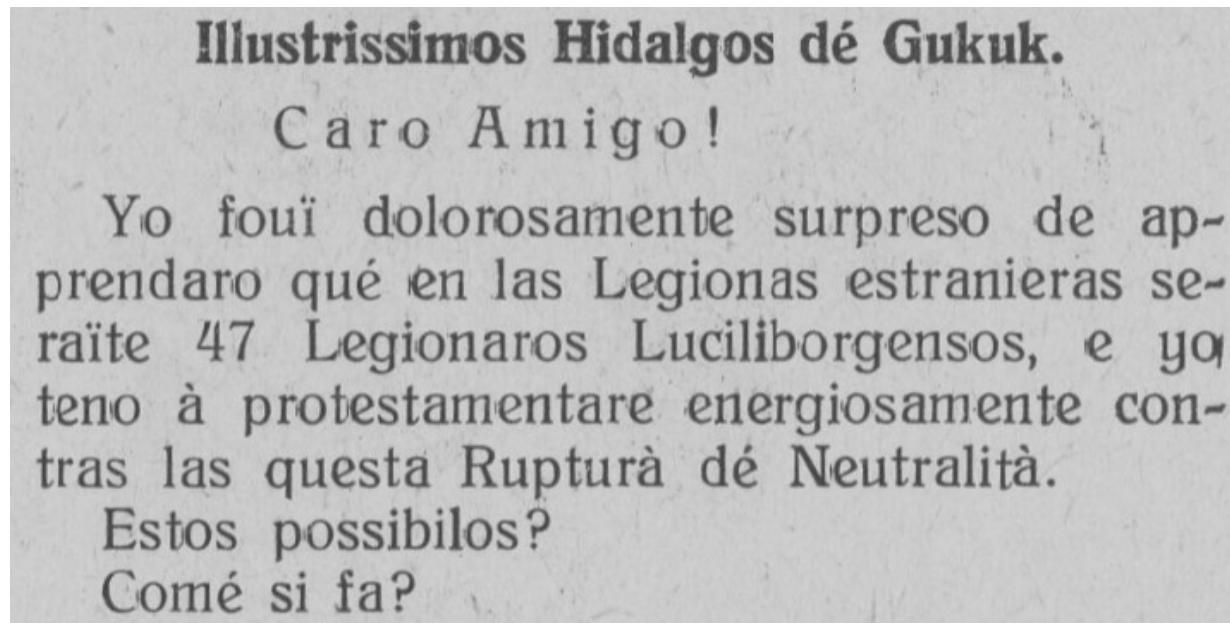
1. Szentestén: Weilerbachan az Institut Heliar-ban, az ottani magyar menekültek székhelyén lesz az éjféli mise.
2. Karácsony első napján: Luxemburgban, a Vereinshaus-ban, a katedrálissal szemben, délelött fél 10 órakor;
Differdingban, Korhaz, déledött 11 órakor ünnepi szentmise.

Minden luxemburgi magyart szereettel várunk.

Finds: invented Language

Satiric article, “fake” Spanish

1925 <https://persist.lu/ark:70795/2747qp1fh/pages/2/articles/DTL92>





Questions?

<https://eluxemburgensia.lu>

Yves Maurer
National Library of Luxembourg

yves.maurer@bnl.etat.lu

@yvesmaurer

@ymaurer@mastodon.top