



Apprentissage Automatique : ???

L. Grobol (MoDyCo, Université Paris Nanterre)

M2 Plurital

Nanterre, France, 2024-10-16

Failure modes

Specification gaming

From Krakovna (2018)

Specification gaming

From Krakovna (2018)

- A boat race agent 'goes in a circle hitting the same targets instead of finishing the race.'

Specification gaming

From Krakovna (2018)

- A **boat race agent** 'goes in a circle hitting the same targets instead of finishing the race.'
- 'A robotic arm trained to slide a block to a target position on a table achieves the goal by moving the table itself.'

Specification gaming

From Krakovna (2018)

- A boat race agent 'goes in a circle hitting the same targets instead of finishing the race.'
- 'A robotic arm trained to slide a block to a target position on a table achieves the goal by moving the table itself.'
- '[Tetris] Agent pauses the game indefinitely to avoid losing.'

Specification gaming

From Krakovna (2018)

- A boat race agent 'goes in a circle hitting the same targets instead of finishing the race.'
- 'A robotic arm trained to slide a block to a target position on a table achieves the goal by moving the table itself.'
- '[Tetris] Agent pauses the game indefinitely to avoid losing.'
- 'Deep learning model to detect pneumonia in chest x-rays works out which x-ray machine was used to take the picture; that, in turn, is predictive of whether the image contains signs of pneumonia, because certain x-ray machines (and hospital sites) are used for sicker patients.'

Specification gaming

From Krakovna (2018)

- A boat race agent 'goes in a circle hitting the same targets instead of finishing the race.'
- 'A robotic arm trained to slide a block to a target position on a table achieves the goal by moving the table itself.'
- '[Tetris] Agent pauses the game indefinitely to avoid losing.'
- 'Deep learning model to detect pneumonia in chest x-rays works out which x-ray machine was used to take the picture; that, in turn, is predictive of whether the image contains signs of pneumonia, because certain x-ray machines (and hospital sites) are used for sicker patients.'
- 'Agent kills itself at the end of level 1 to avoid losing in level 2'

Pourquoi tant de haine

Rappelez-vous du cours d'introduction :

Pourquoi tant de haine

Rappelez-vous du cours d'introduction :

L'apprentissage automatique n'a quasiment rien à voir avec l'apprentissage humain.

Pourquoi tant de haine

Rappelez-vous du cours d'introduction :

L'apprentissage automatique n'a quasiment rien à voir avec l'apprentissage humain.

Ce n'est donc pas que les modèles ici sont malicieux ou quelque autre qualité humaine. Il existe simplement des **régularités** dans les données (dans des cas assez particuliers ici), qui sont suffisamment évidentes pour être repérées et utilisées.

Pourquoi tant de haine

Rappelez-vous du cours d'introduction :

L'apprentissage automatique n'a quasiment rien à voir avec l'apprentissage humain.

Ce n'est donc pas que les modèles ici sont malicieux ou quelque autre qualité humaine. Il existe simplement des **régularités** dans les données (dans des cas assez particuliers ici), qui sont suffisamment évidentes pour être repérées et utilisées.

Pour un·e humain·e, la différence entre exploiter un bug et effectuer légitimement la tâche est claire. Pour un algo d'apprentissage, **c'est exactement la même chose.**

Pourquoi tant de haine

Rappelez-vous du cours d'introduction :

L'apprentissage automatique n'a quasiment rien à voir avec l'apprentissage humain.

Ce n'est donc pas que les modèles ici sont malicieux ou quelque autre qualité humaine. Il existe simplement des **régularités** dans les données (dans des cas assez particuliers ici), qui sont suffisamment évidentes pour être repérées et utilisées.

Pour un·e humain·e, la différence entre exploiter un bug et effectuer légitimement la tâche est claire. Pour un algo d'apprentissage, **c'est exactement la même chose**.

Par contre l'histoire des tanks a bien l'air apocryphe, voir Branwen (2019)

Démo : un modèle de détection d'opinions

Diversity ?

'Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ'

'La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.'

Apertium *'La langue si elle fait avec elle une personne elle l'est un monde s'il vit et mon effort dans lui.'* (Tyers (2010), rule based)

m2m100 *'C'est le cas d'un homme qui a laissé le coucher, et qui a laissé le coucher.'* (Fan et al. (2021), multiling NMT)

'Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ'

'La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.'

OPUS-celtic 'The language if she has a man who lives a world and has my straight.'

OPUS-elg 'This language contains one or more people who have their own world!'

So what's happening?

OPUS (Tiedemann 2012) is a meta parallel corpus with covering for ≈ 400 languages.

For Breton, it includes most of the publicly available data:

br-fr 1.6 Msentences

br-en 1.1 Msentences

OPUS (Tiedemann 2012) is a meta parallel corpus with covering for ≈ 400 languages.

For Breton, it includes most of the publicly available data:

br-fr 1.6 Msentences

br-en 1.1 Msentences

BUT

Something's wrong with OPUS

Most of the Breton data in OPUS comes from

- WikiMatrix (Schwenk et al. 2021a)
- CCMatrix (Schwenk et al. 2021b)
- CCAIined (El-Kishky et al. 2020)

all built from parallel corpus mining:

Something's wrong with OPUS

Most of the Breton data in OPUS comes from

- WikiMatrix (Schwenk et al. 2021a)
- CCMatrix (Schwenk et al. 2021b)
- CCAIined (El-Kishky et al. 2020)

all built from parallel corpus mining:

1. Train a 'language-agnostic' vector sentence representation model e.g. LASER, (Artetxe and Schwenk 2019)
2. Look in monolingual corpora for sentences in different languages but with similar representations.

Something's wrong with OPUS

Most of the Breton data in OPUS comes from

- WikiMatrix (Schwenk et al. 2021a)
- CCMatrix (Schwenk et al. 2021b)
- CCAIined (El-Kishky et al. 2020)

all built from parallel corpus mining:

1. Train a 'language-agnostic' vector sentence representation model e.g. LASER, (Artetxe and Schwenk 2019)
2. Look in monolingual corpora for sentences in different languages but with similar representations.

Only one corpus is actually made of human translations: OPAB (Ofis Publik ar Brezhoneg, Tyers (2009)).

Something's wrong with OPUS

But: the embeddings for Breton are **poorly aligned**.

- Artetxe and Schwenk (2019) report an **error rate of 85 %!**
- Not a surprise! It's trained using the OpenSubtitles Corpus (Lison and Tiedemann 2016):

Something's wrong with OPUS

But: the embeddings for Breton are **poorly aligned**.

- Artetxe and Schwenk (2019) report an **error rate of 85 %!**
- Not a surprise! It's trained using the OpenSubtitles Corpus (Lison and Tiedemann 2016):
 - Small for Breton-* pairs.

Something's wrong with OPUS

But: the embeddings for Breton are **poorly aligned**.

- Artetxe and Schwenk (2019) report an **error rate of 85 %!**
- Not a surprise! It's trained using the OpenSubtitles Corpus (Lison and Tiedemann 2016):
 - Small for Breton-* pairs.
 - Poorly aligned.

What's wrong with OPUS

- *'Super-harozed.'* vs. 'Vous êtes des héros. Des Super héros.'
- *'C'hoant 'moa goût petra 'oa c'hoarvezet.'* vs. *'J'aimais Tony. Je voulais savoir ce qui s'est passé.'*
- *'Me ivez !'* vs. 'What are the chances?'
- *'Kae da sutal, Pakistan brein !'* vs. 'You're sorry?'

Something's wrong with OPUS

But: the embeddings for Breton are **poorly aligned**.

- Artetxe and Schwenk (2019) report an **error rate of 85 %!**
- Not a surprise! It's trained using the OpenSubtitles Corpus (Lison and Tiedemann 2016):
 - Small for Breton-* pairs.
 - Poorly aligned.

Something's wrong with OPUS

But: the embeddings for Breton are **poorly aligned**.

- Artetxe and Schwenk (2019) report an **error rate of 85 %!**
- Not a surprise! It's trained using the OpenSubtitles Corpus (Lison and Tiedemann 2016):
 - Small for Breton-* pairs.
 - Poorly aligned.

No one has actually checked it.

No speaker of Breton has ever been involved in the development of either datasets or models.

Diversity ?

Other crimes

'Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ'

'La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.'

GPT-3.5 *'La langue qu'elle parle est celle d'une personne qui a en elle un monde où elle vit et lutte.'*

'La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.'

'Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ.'

GPT-3.5 *'An teunga a implij ur vro eo ur bed en e ober a blij ha emdroadur.'*

'La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.'

'Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ.'

GPT-3.5 *'An teunga a implij ur vro eo ur bed en e ober a blij ha emdroadur.'*

GPT-3.5 *'Yezh mae eun an hini a labour, ur bed e vev ha ober a raio.'*

'La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.'

'Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ.'

GPT-3.5 *'An teunga a implij ur vro eo ur bed en e ober a blij ha emdroadur.'*

GPT-3.5 *'Yezh mae eun an hini a labour, ur bed e vev ha ober a raio.'*

GPT-3.5 *'Yezh ar re a labour a zo bed ma vev ha ma c'hevrioù.'*

'La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.'

'Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ.'

GPT-3.5 *'An teunga a implij ur vro eo ur bed en e ober a blij ha emdroadur.'*

GPT-3.5 *'Yezh mae eun an hini a labour, ur bed e vev ha ober a raio.'*

GPT-3.5 *'Yezh ar re a labour a zo bed ma vev ha ma c'hevrioù.'*

GPT-3.5 ...

DEGEMER MAT

BIENVENUE

Petra eo ho koulenn ? | Quelle est votre question ?

Gallout a rit skrivañ ur gemennadenn e brezhoneg pe e galleg, ha respont a vo deoc'h dre ziouer e brezhoneg dre zegas. E vo kinniget deoc'h e-barzh krochennege ar frazennoù e galleg. Bez' ez eus plijadur!

Vous pouvez soumettre un message en breton ou en français, notre outil vous répondra automatiquement en breton. La traduction en français vous sera proposée entre parenthèses. Amusez-vous bien !

ChatGPT.bzh : l'intelligence artificielle qui parle breton !

Bienvenue sur notre nouveau site de chatbot en breton ! ChatGPT.bzh est le [ChatGPT](#) breton. [Si vous cherchez à apprendre la langue bretonne ou à améliorer votre niveau, vous êtes au bon endroit.](#) Notre chatbot utilise l'intelligence artificielle pour vous fournir des réponses en breton sur une variété de sujets, allant de la [culture bretonne](#) au sens large, du [vocabulaire](#), à la grammaire et à la prononciation.

Notre chatbot en breton est un outil innovant qui vous permet de pratiquer la langue bretonne en ligne, de manière facile et interactive. En communiquant avec notre chatbot en breton, vous aurez l'opportunité de découvrir une langue riche et fascinante, tout en améliorant votre compréhension et votre expression écrite et orale.

Que vous soyez débutant ou avancé, notre chatbot en breton peut vous aider à progresser. Essayez notre chatbot en breton dès maintenant et découvrez l'expérience unique que nous avons créée pour vous. Notre chatbot est disponible 24h/24, 7j/7 pour répondre à toutes vos questions en breton.

N'hésitez pas à commencer à discuter avec lui dès maintenant et à apprendre la langue bretonne de manière amusante et interactive !

kregiñ ar c'hat

Commencez le chat !

DEGEMER MAT

BIENVENUE

Petra eo ho koulenn ? | Quelle est votre question ?

Gallout a rit se lier d'us gennadennoù brezhoneg pe e galleg, ha respont a vo deoc'h dre z'usur e brezhoneg dre zegas. E vo hinnet deoc'h e-barzh krochenneg ar frazennoù e galleg. Bez' ez eus plijadur!

Vous pouvez soumettre un message en breton ou en français, notre outil vous répondra automatiquement en breton. La traduction en français vous sera proposée entre parenthèses. Amusez-vous bien !

ChatGPT.bzh : l'intelligence artificielle qui parle breton !

Bienvenue sur notre nouveau site de chatbot en breton ! ChatGPT.bzh est le [ChatGPT](#) breton. [Si vous cherchez à apprendre la langue bretonne ou à améliorer votre niveau, vous êtes au bon endroit.](#) Notre chatbot utilise l'intelligence artificielle pour vous fournir des réponses en breton sur une variété de sujets, allant de la [culture bretonne](#) au sens large, du [vocabulaire](#), à la grammaire et à la prononciation.

Notre chatbot en breton est un outil innovant qui vous permet de pratiquer la langue bretonne en ligne, de manière facile et interactive. En communiquant avec notre chatbot en breton, vous aurez l'opportunité de découvrir une langue riche et fascinante, tout en améliorant votre compréhension et votre expression écrite et orale.

Que vous soyez débutant ou avancé, notre chatbot en breton peut vous aider à progresser. Essayez notre chatbot en breton dès maintenant et découvrez l'expérience unique que nous avons créée pour vous. Notre chatbot est disponible 24h/24, 7j/7 pour répondre à toutes vos questions en breton.

N'hésitez pas à commencer à discuter avec lui dès maintenant et à apprendre la langue bretonne de manière amusante et interactive !

kregiñ ar c'hat

Commencez le chat !



Diversity ?

With a little help from field linguistics



Atlas Rannyezhoù ar
BREzhoneg : Sintaks

Navigation

Accueil
Utiliser ce site

Grammaire bretonne

Introduction
Prononciation
Composition de mot
Le nom
La préposition
Le verbe
La phrase
Parler en contexte
-> Article au hasard

Linguistique formelle

Introduction
Architecture
Phonologie
Morphologie
Principaux constituants
Syntaxe de la phrase
Structure
informationnelle
Sémantique

La lénition

[Page](#) [Discussion](#)

[Voir le texte source](#) [Historique](#)

(Redirigé depuis [Mutation douce](#))

La lénition, *mutation consonantique dite 'adoucissante'*, notée sur ce site par un exposant 1, est déclenchée par:

le rannig *a*

les déterminants possessifs *da* (2SG) et *e* (3SGM)

les pronoms objet proclitiques *da* (2SG) et *e* (3SGM)

les prépositions *da*, *a*, *dre*, et parfois après *dindan*, *diwar*, *pe*, *war*

les *cardinaux* *daou* et *div* (2)

les *particules négatives* *ne* et *na*

le réflexif *en em*

l'équivalent du gérondif: *en ur*,

le *quantifieur* *holl*

la préposition *eme*

la composition morphologique, avec *gwall-*, *hanter-*, et de nombreux autres préfixes.

un mot féminin singulier, ou masculin pluriel de personnes qui n'ont pas leur pluriel en *-où* sur un adjectif épithète ou un nom en *apposition*.

un nom propre sur son adjectif en contexte *hypocoristique* (*Per vihan*)

Consonne initiale mutable:	K	T	P	G	Gw	D	B	M
1.	G	D	B	C'h	W	Z	V	V
1 a.	G	D	B	C'h	W	-	V	V
1 b.	-	-	-	C'h	W	Z	V	V

Cas 1a: après l'article défini *ar* et *an*, et l'article indéfini *ur* et *un* pour les noms féminins singuliers et les noms masculins pluriels de personnes qui n'ont pas leur pluriel en *-où*.

Cas 1b: adjectifs épithètes et noms en apposition après les substantifs se terminant par autre chose que L, M, N, R, V ou une voyelle.

Plus

[Pages liées](#)
[Suivi des pages liées](#)
[Version imprimable](#)
[Lien permanent](#)
[Informations sur la page](#)
[Journaux de la page](#)

Liste des catégories

[Articles](#)
[Phonologie](#)

Atlas Rannyezhoù ar **BRE**zhoneg: Sintaks (Jouitteau 2009–2024)

- A public, collaborative and evolutive **research notebook** for fundamental research in formal linguistics.
- A descriptive **grammar for the speaking community**.
- Including **illustrative data**
 - News
 - Cultural/artistic productions
 - Social media content
 - **Elicitation** data collected in linguistic **fieldwork**

Breton sentences and their translations extracted from ARBRES's interlinear glosses:

- (1) *eur mell gwezenn glas he deliou*
a big tree.SG green his² leaf.PL
“ a big tree with green leaves ”

After extraction and deduplication we got **5192 sentences**

After extraction and deduplication we got **5192 sentences**

- An order of magnitude less than the *Ofis Publik ar Brezhoneg* parallel corpus.

After extraction and deduplication we got 5192 sentences

→ An order of magnitude less than the *Ofis Publik ar Brezhoneg* parallel corpus.

So is it all for nothing?

After extraction and deduplication we got 5192 sentences

→ An order of magnitude less than the *Ofis Publik ar Brezhoneg* parallel corpus.

So is it all for nothing?

No!

For a more precise comparison, we evaluate:

Apertium Mostly rule-based, developed for Breton by Tyers (2010) on the OPAB corpus.

m2m100-418m out-of-the-box.

+OPUS with continued pretraining on OPUS data (with heuristic filtering for quality, so mostly OPAB)

+ARBRES with both OPUS and ARBRES Kenstur.

Evaluation on a (sadly) for now private dataset from OPAB.

Results

Model	BLEU	ChrF++	TER
Apertium	24.15	50.23	63.93
m2m100-418M	0.58	11.85	114.49
+OPAB	30.01	50.16	55.37
+ARBRES	37.68	56.99	48.65
+Korpus Nevez	40.37	59.14	44.10

Evaluation results on the OPAB test dataset.

Results

Model	BLEU	ChrF++	TER
Apertium	24.15	50.23	63.93
m2m100-418M	0.58	11.85	114.49
+OPAB	30.01	50.16	55.37
+ARBRES	37.68	56.99	48.65
+Korpus Nevez	40.37	59.14	44.10

Evaluation results on the OPAB test dataset.

- Apertium fares surprisingly well
- Despite the limited size of ARBRES, the gain is important.

Results

Model	BLEU	ChrF++	TER
Apertium	24.15	50.23	63.93
m2m100-418M	0.58	11.85	114.49
+OPAB	30.01	50.16	55.37
+ARBRES	37.68	56.99	48.65
+Korpus Nevez	40.37	59.14	44.10

Evaluation results on the OPAB test dataset.

- Apertium fares surprisingly well
- Despite the limited size of ARBRES, the gain is important.
- ALSO LOOK AT THAT LAST LINE OMG

But what does it look like?

'Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ'

'La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.'

m2m100 *'C'est le cas d'un homme qui a laissé le coucher, et qui a laissé le coucher.'*

+OPUS *'La langue dans laquelle elle fait un homme est un monde dans lequel elle vit et s'efforce.'*

+OPUS+ARBRES *'La langue dans laquelle un homme parle est un monde dans lequel il vit et s'efforce.'*

Diversity ?

What now?

Why?????

For companies and research institutions, there is a real incentive to handle minority languages.

Why?????

For companies and research institutions, there is a real incentive to **claim to** handle minority languages.

Why?????

For companies and research institutions, there is a real incentive to claim to handle minority languages.

But very little to actually do it well: doing poorly has very little consequences.

Why?????

For companies and research institutions, there is a real incentive to claim to handle minority languages.

But very little to actually do it well: doing poorly has very little consequences.

This leads to a form of **diversity washing**, with plausible-ish deniability.

Why?????

For companies and research institutions, there is a real incentive to claim to handle minority languages.

But very little to actually do it well: doing poorly has very little consequences.

This leads to a form of **diversity washing**, with plausible-ish deniability.

Whenever there is at least some data available, just use it. No need for any other commitment: **the subaltern cannot speak** (Chakravorty Spivak 1988).

Consequences

By increasing severity:

- Proliferation of erroneous language, potentially contaminating uses of even native speakers.

Consequences

By increasing severity:

- Proliferation of erroneous language, potentially contaminating uses of even native speakers.
- Hiding the lack of actual language technologies, leading to lack of investments by public actors.

Consequences

By increasing severity:

- Proliferation of erroneous language, potentially contaminating uses of even native speakers.
- Hiding the lack of actual language technologies, leading to lack of investments by public actors.
- Linguistic conflict between NLP tools and actual speakers, furthering their dispossession and silencing.

Consequences

By increasing severity:

- Proliferation of erroneous language, potentially contaminating uses of even native speakers.
- Hiding the lack of actual language technologies, leading to lack of investments by public actors.
- Linguistic conflict between NLP tools and actual speakers, furthering their dispossession and silencing.

Worst of all: producers of such NLP tools have a direct economic incentive in silencing the linguistic communities!

Consequences

By increasing severity:

- Proliferation of erroneous language, potentially contaminating uses of even native speakers.
- Hiding the lack of actual language technologies, leading to lack of investments by public actors.
- Linguistic conflict between NLP tools and actual speakers, furthering their dispossession and silencing.

Worst of all: producers of such NLP tools have a direct economic incentive in silencing the linguistic communities!

In many cases, these minority linguistic communities are already in difficult positions re: literacy in English, higher education, competences in CS and linguistics, and access to media.

On a basic level:

- Take claims of massive multilingualism with a big grain of salt.

On a basic level:

- Take claims of massive multilingualism with a big grain of salt.
- Be very careful with uncontrolled parallel data mining.

On a basic level:

- Take claims of massive multilingualism with a big grain of salt.
- Be very careful with uncontrolled parallel data mining.
- **Curating** a **focused** dataset is worth the effort.

More importantly:

- NLP without **language experts** makes very little sense.

More importantly:

- NLP without **language experts** makes very little sense.
- NLP **by** language experts is a lot better. In all respects.

More importantly:

- NLP without **language experts** makes very little sense.
- NLP **by** language experts is a lot better. In all respects.
- **Linguists** and their data are far too precious to ignore.

What to do

Most importantly:

- Stop thinking of NLP experts, language experts and linguistic communities as separate entities.

Most importantly:

- Stop thinking of NLP experts, language experts and linguistic communities as separate entities.
- Stop thinking of NLP as a collect data → train → evaluate pipeline with separate actors.

What to do

Most importantly:

- Stop thinking of NLP experts, language experts and linguistic communities as separate entities.
- Stop thinking of NLP as a collect data → train → evaluate pipeline with separate actors.
- **Listening** to linguistic communities is good, but it's not enough:

Most importantly:

- Stop thinking of NLP experts, language experts and linguistic communities as separate entities.
- Stop thinking of NLP as a collect data → train → evaluate pipeline with separate actors.
- **Listening** to linguistic communities is good, but it's not enough:
 - Teach.

Most importantly:

- Stop thinking of NLP experts, language experts and linguistic communities as separate entities.
- Stop thinking of NLP as a collect data → train → evaluate pipeline with separate actors.
- **Listening** to linguistic communities is good, but it's not enough:
 - Teach.
 - Support.

What to do

Most importantly:

- Stop thinking of NLP experts, language experts and linguistic communities as separate entities.
- Stop thinking of NLP as a collect data → train → evaluate pipeline with separate actors.
- **Listening** to linguistic communities is good, but it's not enough:
 - Teach.
 - Support.
 - **Don't be an outsider.**

Mental illusions

https://www.youtube.com/watch?v=wp8ebj_yRI4

Appendix

References i

- Artetxe, Mikel and Holger Schwenk (1st Sept. 2019). **'Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond'**.
In: *Transactions of the Association for Computational Linguistics* 7.
URL: https://doi.org/10.1162/tacl_a_00288.
- Branwen, Gwern (7th Feb. 2019). ***The Neural Net Tank Urban Legend***.
URL: <https://www.gwern.net/Tanks> (visited on 21/02/2019).
- Chakravorty Spivak, Gayatri (1988). **'Can the Subaltern Speak'**.
In: *Marxism and the Interpretation of Culture*. University of Illinois Press.
URL: <https://jan.ucc.nau.edu/~sj6/Spivak%20CanTheSubalternSpeak.pdf>.
- Fan, Angela et al. (1st Jan. 2021). **'Beyond English-Centric Multilingual Machine Translation'**.
In: *The Journal of Machine Learning Research* 22.1.
URL: <https://dl.acm.org/doi/abs/10.5555/3546258.3546365>.

References ii

Jouitteau, Mélanie (2009–2024). **ARBRES, Wikigrammaire Des Dialectes Du Breton et Centre de Ressources Pour Son Étude Linguistique Formelle.**

URL: <http://arbres.iker.cnrs.fr>.

El-Kishky, Ahmed, Vishrav Chaudhary, Francisco Guzmán and Philipp Koehn (Nov. 2020).

‘CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs’.

In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Online: Association for Computational Linguistics.

URL: <https://aclanthology.org/2020.emnlp-main.480>.

Krakovna, Victoria (2nd Apr. 2018).

Specification Gaming Examples in AI - Master List.

URL: [https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-](https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bj0xCG84dAg/pubhtml)

[32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bj0xCG84dAg/pubhtml](https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bj0xCG84dAg/pubhtml) (visited on 21/02/2019).

References iii

- Lison, Pierre and Jörg Tiedemann (May 2016). **'OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles'**. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. LREC 2016. Portorož, Slovenia: European Language Resources Association (ELRA). URL: <https://aclanthology.org/L16-1147>.
- Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong and Francisco Guzmán (Apr. 2021a). **'WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia'**. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. EACL 2021. Online: Association for Computational Linguistics. URL: <https://aclanthology.org/2021.eacl-main.115>.

References iv

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin and Angela Fan (Aug. 2021b).

'CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web'.

In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL-IJCNLP 2021.

Online: Association for Computational Linguistics.

URL: <https://aclanthology.org/2021.acl-long.507>.

Tiedemann, Jörg (May 2012). **'Parallel Data, Tools and Interfaces in OPUS'.**

In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. LREC 2012.

Istanbul, Turkey: European Language Resources Association (ELRA).

URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

References v

- Tyers, Francis M. (14th May 2009). **'Rule-Based Augmentation of Training Data in Breton-French Statistical Machine Translation'**. In: *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*. EAMT 2009 (Barcelona, España). European Association for Machine Translation. URL: <https://aclanthology.org/2009.eamt-1.29>.
- Tyers, Francis M. (27th May 2010). **'Rule-Based Breton to French Machine Translation'**. In: *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. EAMT 2010. Saint Raphaël, France: European Association for Machine Translation. URL: <https://aclanthology.org/2010.eamt-1.13>.



This document is distributed under the terms of the Creative Commons
Attribution 4.0 International Licence (CC BY 4.0)
(creativecommons.org/licenses/by/4.0)

© 2024, L. Grobol <loic.grobol@gmail.com>

lgrobol.eu