

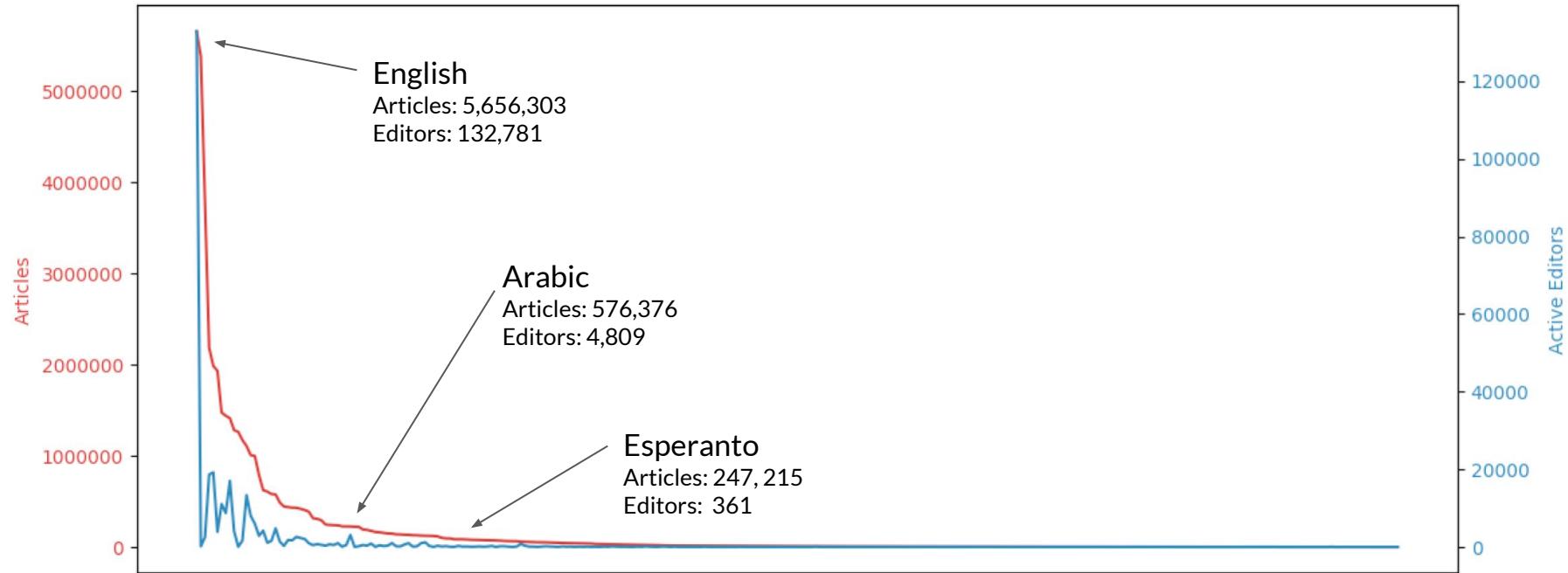
---

# Mind the (Language) Gap: Neural Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders



Lucie-Aimée Kaffee\*, Hady Elsahar\*, Pavlos Vougiouklis\*, Christophe Gravier,  
Frédérique Laforest, Jonathon Hare, and Elena Simperl

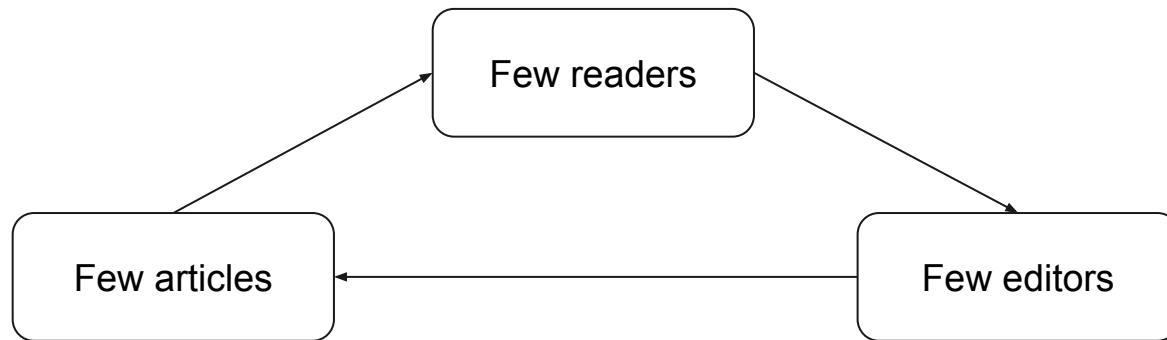
\* the authors contributed equally to this work



Wikipedia is available in 285 languages, but the content is unevenly distributed

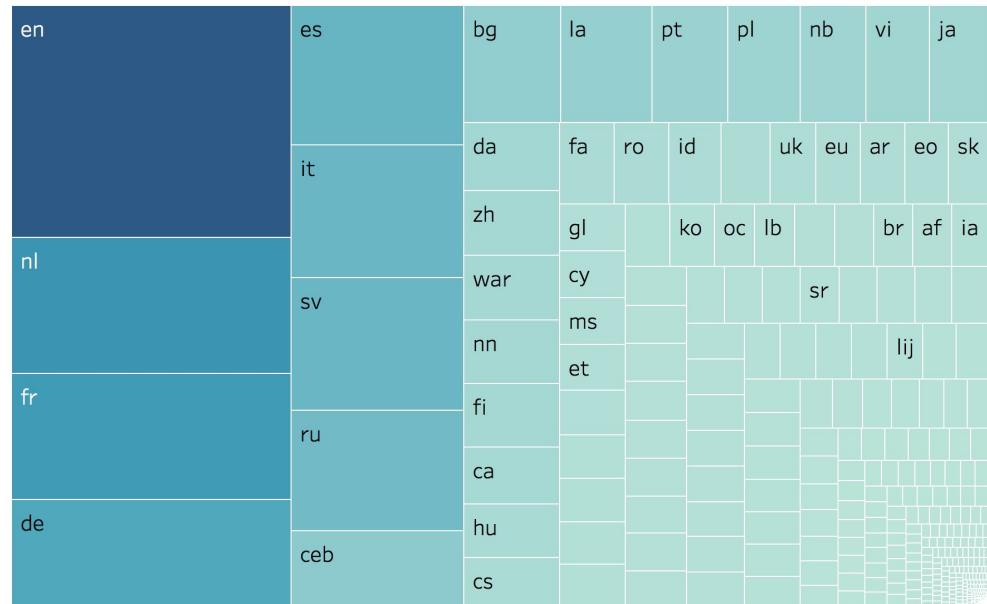
---

## Vicious cycle of lack of information

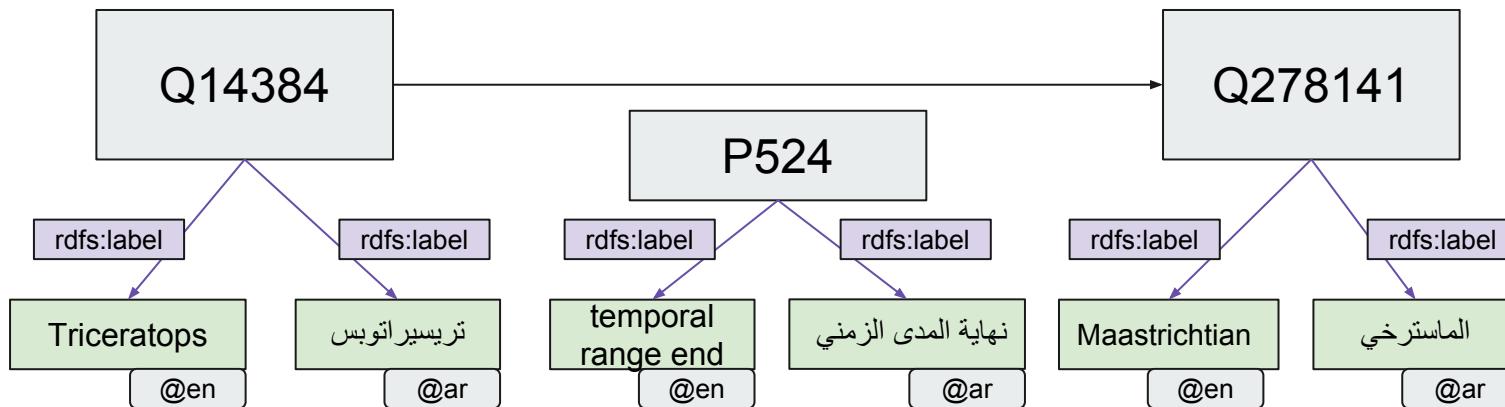


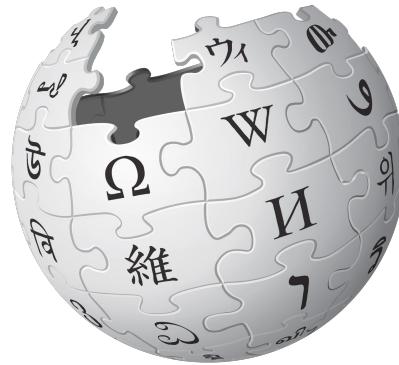
## Wikidata (a source of multilingual structured information )

- Knowledge base maintained and edited by a community of users
- 48,775,926 items
- Each entity can have labels in >400 languages
- Variety of languages well covered (Kaffee et al. 2017)



# Multilinguality in Wikidata





**WIKIPEDIA**  
The Free Encyclopedia

# ArticlePlaceholder

displays Wikidata triples on  
Wikipedia in tabular way

Dynamically generated when data  
changes, not stub articles

Currently deployed on 14  
under-resourced Wikipedias  
(e.g. Gujarati, Haitian Creole, Urdu)

Screenshot of the Haitian Creole Wikipedia article for Triceratops, showing a table of Wikidata triples.

Propriété	Valeur	Source
egzamp nan	taxon fossile	
nom scientifique du taxon	Titanosaurus	auteur taxinomique: Othniel Charles Marsh date de description scientifique: 1890
rang taxinomique	genre	
taxon supérieur	Ceratopidae	
nom vernaculaire	Triceratops	العنكبوت المثلثي (Triceratops prorsus) العنكبوت المثلثي (Triceratops horridus) 3 pieces specimen at the Natural History Museum of Los Angeles
période de disparition	Maastrichtien	
période d'apparition	Maastrichtien	
hauteur	3 Mèt (mez)	
masse	8 500 Kilogram	
longueur	9 Mèt (mez)	
galerie Commons	Triceratops	
catégorie Commons	Triceratops	

Références:

- 1 : bibliothèque du Congrès, 15 janvier 2018
- 2 : sauvegarde de la base de données Freebase, 28 octobre 2013
- 3 : Fossilworks, 13 mai 2015
- 4 : Interim Register of Marine and Nonmarine Genera, 19 avril 2018
- 5 : Encyclopédie de la Vie, 30 octobre 2014
- 6 : Répertoire d'autorité matière encyclopédique et alphabétique unifiée, 15 janvier 2018
- 7 : Système Mondial d'Information sur la Biodiversité, 10 décembre 2018



ویکیپدیا  
آزاد دانش المعرف

## celtic knot

decorative knot

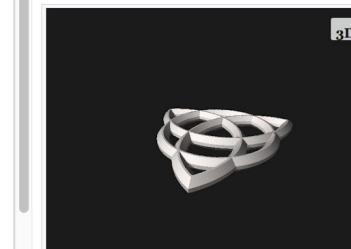
کومنز نگارخانہ

Knots in traditional art

ذیلی درجہ

Islamic interlace patterns  
decorative knot

3D model



کومنز زمرہ

Celtic knots

بیرونی وسائل  
فری میں آئی ذہنی /  
[i]m/ofk\_3/

صلیل اول  
 موضوعات ابواب  
 جستجوی مطالعہ  
 یا مضمون تحریر کریں  
 راست کریں

تمام

ویکیپدیا پر آغاز کریں  
 معاویت  
 دریافت عام  
 حاکمیتیں  
 مطہریہ  
 اپلود تصویر

آلات

خصوصی صفحات  
قابل طبع نسخہ  
 ذیلی آخری

دیگر منصوبے  
 ویکیپدیا العالم

مضمون تحریر کریں

حوالہ

۱. فری میں دینا دھس، 28 اکتوبر 2013ء

دیگر زبانیں

English  
Nederlands  
Português  
Español  
Deutsch

# ArticlePlaceholders “Textual Extension”



VIKIPEDIO  
La libera enciklopedio

Speciala

## celtic knot

decorative knot

Missing descriptions  
(English ones are used)  
Wikidata specific

sub-aro de

komuneja bildaro

The screenshot shows a Wikipedia page for "celtic knot". The page title is "celtic knot". Below the title, there is a box containing the text "decorative knot". An orange rectangle highlights this text. A black arrow points from this highlighted text to the right, where the text "Missing descriptions (English ones are used) Wikidata specific" is displayed. Below the title, there are two other boxes: one containing "sub-aro de" and another containing "komuneja bildaro". On the left side of the page, there is a sidebar with links such as "Cefpaĝo", "Komunuma portalo", "Diskutejo", "Aktualajoj", and "Lastaj ŝanĝoj". The top of the page has a navigation bar with "Speciala" selected.

# ArticlePlaceholders “Textual Extension”

Enriching ArticlePlaceholder with textual summaries generated from Wikidata triples using Neural Language Generation from Structured Knowledge bases

The screenshot shows a Wikipedia article page for "celtic knot". The page has a header with the title "celtic knot" and a sub-page navigation bar with links like "Speciala", "Serĉi tra Vikipe", and "PEDIO enciklopedio". The main content area contains a green-bordered summary box with the text: "Keltaj nodoj, nomitaj lkovellavna, estas diversaj nodoj kaj stiligitaj grafikaj reprezentoj de nodoj uzataj por ornamado, uzitaj vaste en la kelta stilo de Insula arto". Below this, there are two boxes: one for "sub-aro de" (with "Islamic interlace patterns") and another for "komuneja bildaro" (with "Knots in traditional art"). On the right side, there is a sidebar with "Eksteraj r" and "identigilo de F". The left sidebar of the page shows parts of the page structure like "na portalo", "j", "nĝoj", and "soi".

# Enriching ArticlePlaceholder with textual summaries generated from Wikidata triples using Neural Language Generation from Structured Knowledge bases

- More pleasant to readers than tables
- Can serve as a starting point to a wikipedia article

Working on under-resourced languages ( testing on Arabic and Esperanto)

The screenshot shows a Wikipedia article page for the triceratops. At the top, there's a sidebar with various language links and a search bar. The main content area has a placeholder for the first section, which is filled with a summary of taxonomic information. This summary includes:  
- Nom scientifique du taxon: Triceratops (with a note: "auteur taxinomique: Othniel Charles Marsh date de description scientifique: 1890")  
- Rang taxinomique: genre  
- Taxon supérieur: Ceratopidae  
- Nom vernaculaire: Triceratops (with notes: "الثلاثي الرمادي", "الثلاثي", "الثلاثي الرمادي", "Triceratops", "三角龍")  
- Période de disparition: Maastrichtien  
- Période d'apparition: Maastrichtien  
- Hauteur: 3 Mét (mèz)  
- Longueur: 9 Mèt (mèz)  
- Masse: 8 500 Kilogram  
- Galerie Commons: Triceratops (with a note: "catégorie Commons: Triceratops")  
- Catégorie Commons: Triceratops

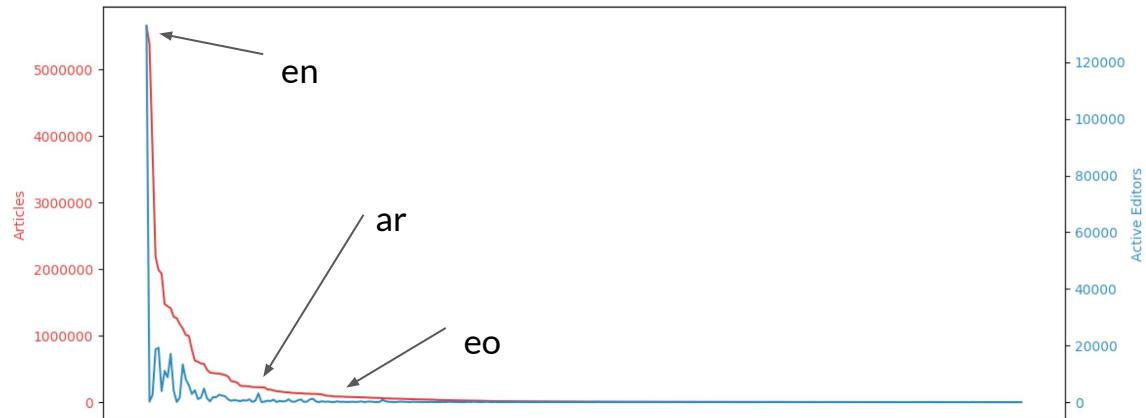
Below the placeholder, there's a "Références" section with a numbered list of sources, and a "Ressources extérieures" section with a list of external identifiers.

# Esperanto

- Esperanto is an artificial language
- Easy to learn
- Engaged Wikipedia community
- A good starting point

# Arabic

- Arabic is the 5th most spoken language in the world
- Content online in Arabic is sparse however





- displays Wikidata information
- manual created templates
- community effort
- can only display as much information as available in Wikidata

celtic knot (Q242009)

Knotenmuster  
decorative knot

Other properties

**sub-aro de** Islamic interlace patterns ornements évoquant des cordes sans extrémités et enchevêtrées  
decorative knot

Classification

Name	Description
Erito	io, kio ekzistas
Objekto	technical term in modern philosophy often used in contrast to the term subject
Concrete object	object with a physical referent
Fizika substanco	substance composed of quantum particle(s)/field(s), such as matter and/or radiation; that of which objects/systems are composed; physical stuff that can be considered concrete (not strictly abstract)
Fizika korpo	singular aggregation of substance(s) such as matter or radiation, with overall properties such as mass, position or momentum
Artefakto	physical object made or shaped by human hand
Mašinlemento	elementary component of a machine
Rapidumskatolo	machine in a power transmission system for controlled application of the power,gearbox,uses gears/gear trains to provide speed/torque conversions from a rotating power source to another device;reduces the higher engine speed to the slower wheel speed
Pivoto	assembly of bodies connected to manage forces and movement
Nodo	method of fastening or securing linear material, such as rope, by tying or interweaving
Decorative knot	
Interlace	decorative element of bands or portions of other motifs looped, braided, and knotted in complex geometric patterns
Islamic interlace patterns	ornements évoquant des cordes sans extrémités et enchevêtrées
Celtic knot	decorative knot

Related media

Komuneja kategorio : Celtic knots  
komuneja bildaro : Knots in traditional art

Free images Google search

External sources

3D model Celtic knot.stl  
Firebase /m/0fk\_3

Wikimedia projects

Big Wikipedia

cs	Pleteneč (ornament)
de	Knotenmuster
en	Celtic knot
es	Nudo celta
nl	Keltische knoop
pl	Wezel celtycki
pt	Nó celta
ru	Кельтский узел

Wikimedia Commons

commons	Category:Celtic knots
---------	-----------------------

Other Wikispaces

eu	Zelta korapilo
gl	Nó celta
no	Keltisk knute

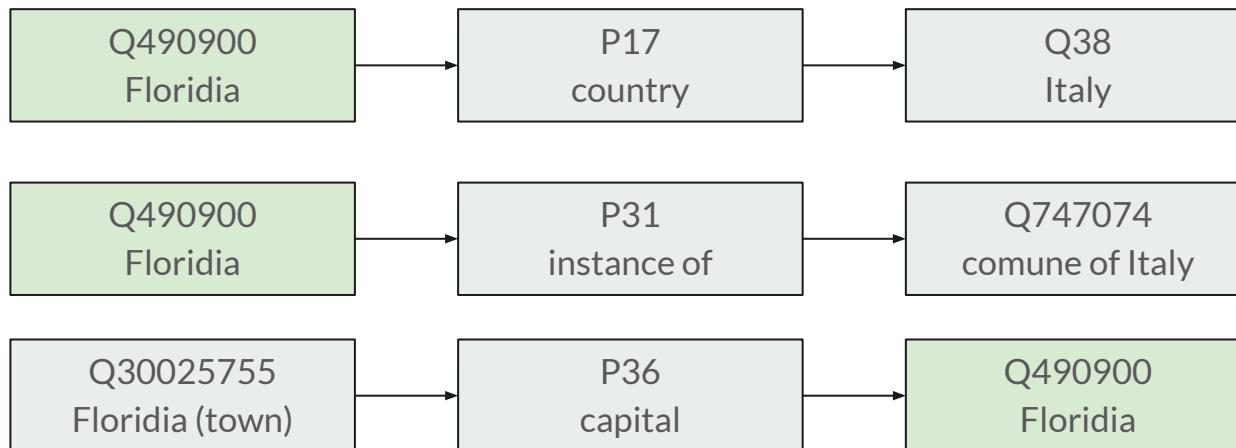
Concept cloud



<https://tools.wmflabs.org/reasonator/?q=Q242009&lang=eo>

---

## Sample Input (from ArticlePlaceholder)

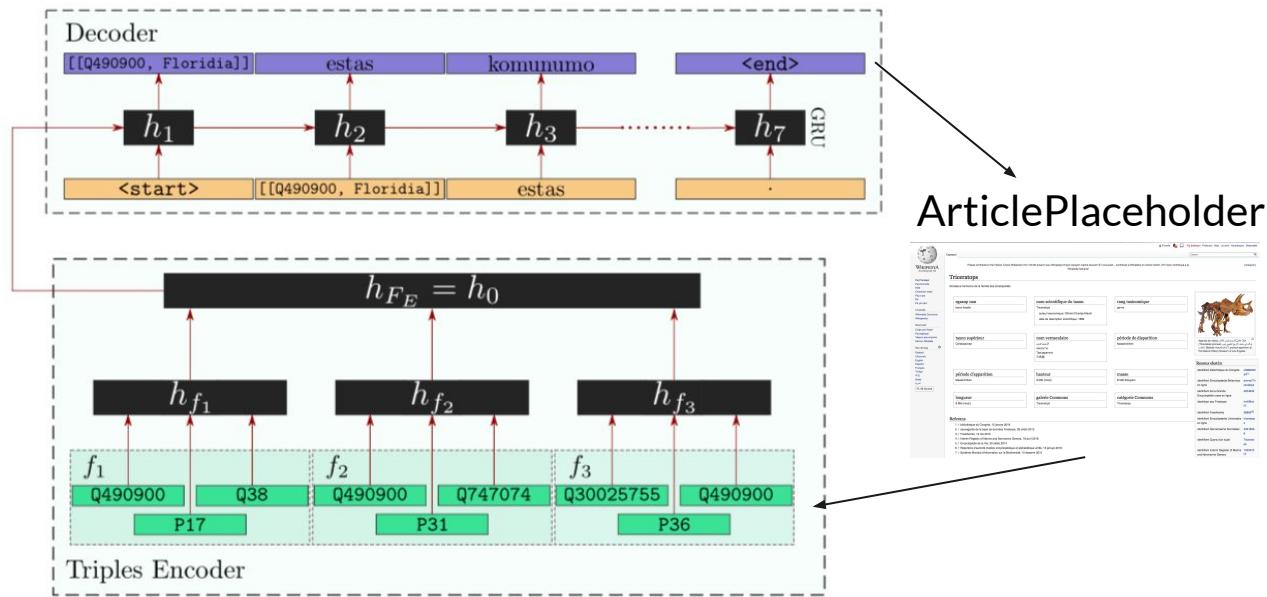


# Neural Text Generation

Feed-forward architecture encodes triples from the ArticlePlaceholder into a vector of fixed dimensionality

RNN-based decoder generates text summaries, one token at a time

Based on [Vougiouklis et al. (2017)]



---

# Dataset Preparation

---

# Dataset Preparation

## Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo (*Calocitta colliei*)** estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseriformaj kiuj loĝas en nordokcidenta Meksiko.

*The black-throated magpie-jay (*Calocitta colliei*) is a strikingly long-tailed magpie-jay of northwestern Mexico.*

---

# Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei



WIKIPEDIA  
The Free Encyclopedia

La **Nigragorĝa pigogarolo (Calocitta colliei)** estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseriformaj kiuj loĝas en nordokcidenta Meksiko.

*The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.*

# Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei



La **Nigragorĝa pigogarolo (*Calocitta colliei*)** estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseriformaj kiuj loĝas en nordokcidenta Meksiko.

*The black-throated magpie-jay (*Calocitta colliei*) is a strikingly long-tailed magpie-jay of northwestern Mexico.*

# Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (*Calocitta colliei*) estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseriformaj kiuj loĝas en nordokcidenta Meksiko.

*The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.*

# Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (**Calocitta colliei**) estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseriformaj kiuj loĝas en nordokcidenta Meksiko.

*The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.*

# Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (**Calocitta colliei**) estas rimarkinda longvosta **pigogarolo** de la familio de Korvedoj kaj ordo de Paseriformaj kiuj loĝas en nordokcidenta Meksiko.

*The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.*

# Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (**Calocitta colliei**) estas rimarkinda longvosta **pigogarolo** de la familio de Korvedoj kaj ordo de Paseriformaj kiuj loĝas en nordokcidenta Meksiko.

*The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.*

Wikimedia's global language fallback chain



## Property Placeholder

- for underserved languages we experience more out of vocabulary words due to limited training data
- to overcome this problem, we introduce property placeholder



# Property Placeholder

Nigragorĝa pigogarolo (Q1586267)

taksonomia nomo (P225)

Calocitta

supera taksono (P171)

Pigogarolo



# Property Placeholder

Nigragorĝa pigogarolo (Q1586267)

taksonomia nomo (P225)

Calocitta

supera taksono (P171)

Pigogarolo

La Nigragorĝa pigogarolo (**Calocitta** colliei) estas rimarkinda longvosta **pigogarolo** de la familio...

# Property Placeholder

Nigragorĝa pigogarolo (Q1586267)

taksonomia nomo (P225)

Calocitta

supera taksono (P171)

Pigogarolo

La Nigragorĝa pigogarolo (**Calocitta** colliei) estas rimarkinda longvosta **pigogarolo** de la familio...



# Property Placeholder

Nigragorĝa pigogarolo (Q1586267)

taksonomia nomo (**P225**)

Calocitta

supera taksono (**P171**)

Pigogarolo

La Nigragorĝa pigogarolo (**Calocitta colliei**) estas rimarkinda longvosta **pigogarolo** de la familio...

La Nigragorĝa pigogarolo ( **[[P225]]** **colliei** ) estas rimarkinda longvosta **[[P171]]** de la familio...

---

# Evaluation

- Automatic Evaluation
- Community Study
  - Readers Evaluation
  - Editors Evaluation

# Generated Examples

---

## Q106693 Group 14 (chemical series):

مجموعة الكربون هي العناصر الموجودة الموجودة في الجدول الدوري للعناصر

Karbongrupo estas elemento en grupo 0 de la perioda tabelo laŭ la IUPAC-sistemo .

The carbon group is a periodic table group consisting of carbon, silicon, germanium, tin, lead, and flerovium.

## Q16885 Thelxinoe (natural satellite):

ثيليكسيون هو قمر طبيعي غير نظامي يتحرك بحركة تراجعية تابع للكوكب المشتري .

Telksino estas neregula satelito de Jupitero , kiu havas retrogradan orbiton .

Thelxinoe (/θɛlk'sɪnəʊ̯ i:/ /θɛlk-SIN-o-ee; Greek: Θελξινόη), also known as Jupiter XLII, is a natural satellite of Jupiter.

---

# Automatic Evaluation

- Baselines
  - Machine Translation → already used in Wikipedia
  - Kneser-Ney → 5-gram language model
  - Template Retrieval → widely used for text generation
- Automatic Evaluation Metrics: BLEU 1 - 4, METEOR, ROUGE

Model	BLEU 4	ROUGE-L	METEOR
MT	9.11	30.51	30.10
KN	0.61	17.09	29.02
KN_ext	13.42	28.52	30.43
TR	25.98	43.58	33.33
TR_ext	32.51	50.57	34.25
Ours	39.20	64.64	45.99
+Placeholders	<b>39.51</b>	<b>64.69</b>	<b>46.17</b>

**Arabic**

Model	BLEU 4	ROUGE-L	METEOR
MT	9.11	30.51	30.1
KN	2.79	36.90	30.74
KN_ext	8.79	44.77	33.71
TR	24.30	45.92	20.46
TR_ext	32.41	57.62	31.04
Ours	34.85	<b>67.02</b>	<b>41.13</b>
+Placeholders	<b>34.95</b>	66.61	40.74

**Esperanto**

Results of the automatic evaluation: Our network outperforms all baselines in Automatic Evaluation

---

## Community Study

- Two 15 days online surveys, aimed at readers and editors in Esperanto and Arabic
- Aiming to test our work with the actual Wikipedia community, outreach on Wikipedia platforms
- Two groups: **readers** and **editors**

---

## Recruitment

- **Readers:** Social media (Reddit Esperanto, Twitter, Facebook)
- **Editors:** Social media, mailing lists and Wikipedia community pages

رجاء المشاركة في استطلاع الرأي الموجود  
على هذا الرابط لدعم دراسة علمية عن ويكيبيديا

<https://isurvey.soton.ac.uk/25272>

Link to our survey featured in the WikiArabia opening remarks



## Reader evaluation

- **Fluency:** Is the text understandable and grammatically correct?
  - ◆ Scores from 0 to 6
- **Appropriateness:** Does the summary ‘feel’ like a Wikipedia article?
- Three different sources:
  - ◆ Generated sentences
  - ◆ Wikipedia
  - ◆ News



# Participation

		Participants	Sentences	Participants Sentences > 50 %	Avg Sentence/ Participant	Total Annotations
<b>Arabic</b>	Fluency	27	60	5	15.03	406
	Approp.	27	60	5	14.78	399
<b>Esperanto</b>	Fluency	27	60	3	8.7	235
	Approp.	27	60	3	8.63	233

## Instructions

من فضلك قم بتقييم جودة النص على مقياس من صفر 0 إلى ستة 6

## Generated Summary

خمساً كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة) ، ويكون على شكل بلورات بيضاء.

قم بتقييم جودة هذا النص:

- 0
- 1
- 2
- 3
- 4
- 5
- 6

## Scores

## Instructions

قيم إذا كنت تعتقد أن هذا النص من الممكن أن يكون مقتبس من الموسوعة الحرة ويكيبيديا العربية أم لا.

لا تعتمد على أي مصادر خارجية لمعرفة الإجابة (مثل محرك بحث جوجل أو ويكيبيديا)

## Generated Summary

خمساً كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة) ، ويكون على شكل بلورات بيضاء.

هل تعتقد أن الجملة السابقة بإمكانها أن تستند إلى جملة في مقال من مقالات ويكيبيديا العربية؟

		Fluency		Appropriateness	
		Mean	SD	Part of Wikipedia	
Arabic	Ours	4.7	1.2	77%	
	Wikipedia	4.6	0.9	74%	
	News	5.3	0.4	35%	
Esper.	Ours	4.5	1.5	69%	
	Wikipedia	4.9	1.2	84%	
	News	4.2	1.2	52%	

Results of the reader study: We generate sentences of comparable fluency, that “feel” like Wikipedia sentences

---

## Editor evaluation

- Editors were asked to edit the article starting from our summary and the corresponding triples (2-3 sentences)
- How much of the text was reused?

# How much of the text was reused by editors?

## Greedy String-Tiling (GST) (0 → 1):

- Mainly used for Plagiarism detection
- Detects whole block moves unlike Levenshtein distance.
- minimum match length (mml) factor to ignore copying of small subsequences
- Dividing results into:
  - WD (wholly derived)
  - PD (partially derived)
  - ND (non-derived)

$$gstscore(S, D) = \frac{\sum_{t_i \in T} |t_i|}{|S|}$$

Length of the Longest common tiles (> mml) between source and edited text

Length of the source text



# Participation

		Participants	Sentences	Participants Sentences > 50 %	Avg Sentence/ Participant	Total Annotations
<b>Arabic</b>	Editors	7	30	2	4	33
<b>Esperanto</b>	Editors	8	30	2	4.75	38

## Instructions

من فضلك قم بكتابة فقرة من الممكن أن تستخدم كأول فقرة في صفحة ويكيبيديا الخاصة بهذا الموضوع باستخدام المعلومات المعطاة لك فقط.

خمساً كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة) ، ويكون على شكل بلورات بيضاء.

كلوريد الزرنيخ الخامس -- حالة خاصة من -- مركب كيميائي

كلوريد الزرنيخ الخامس -- الصيغة الكيميائية -- AtCl<sub>2085</sub>

كلوريد الزرنيخ الخامس -- يتكون من -- زرنيخ

كلوريد الزرنيخ الخامس -- يتكون من -- كلور

كلوريد الزرنيخ الخامس -- مواصفات الإدخال النصي المبسط للجزئيات -- [Cl-].[Cl-].[Cl-].[Cl-].[Cl-].[At+5]

كلوريد الزرنيخ الخامس -- تصنيف كومنز -- Artenic

## Generated Summary

Wikidata triples

Editing field

	Category	Examples	%
Arabic	WD	خامسي كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة)، ويكون على شكل بلورات بيضاء. B A	
	PD	خامسي كلوريد الزرنيخ هو مركب كيميائي له الصيغة (AtClu2085)، ويكون على شكل بلورات بيضاء. B A	45.45%
	ND	بيتش ياتوم أوهابيو (بالإنجليزية (كلمة ناقصة) Ohio (هي منطقة سكنية تقع في الولايات المتحدة في (كلمة ناقصة). C D بيتش ياتوم (بالإنجليزية: Beach Batom) هي قرية تقع في الولايات المتحدة الأمريكية في برووك كاونتي. C D E	33.33%
Esperanto	WD	دیر علا هي بلدة تقع في جنوب غرب ایران. F	
	PD	دیر علا، او بیثر، هي قرية أردنية F	21.21%
	ND	Zederik estas komunumo en la nederlanda provinco Zuid-Holland. g Zederik estas komunumo en la nederlanda provinco Zuid-Holland kaj estas ĉirkaŭata de la municipoj Lopik kaj Zederik. g	78.98%
Esperanto	WD	Nova Pádua estas municipio en la brazila subŝtato Suda Rio-Grando, kiu havis (manka nombro) loĝantojn en (jaro). H	
	PD	Nova Pádua estas municipio en la brazila subŝtato Suda Rio-Grando. H	15.79%
	ND	Ibiúna estas municipio de la brazila subštato San-Paúlio, kiu taksis (manka nombro) enloĝantojn en (jaro). I K L Ibiúna estas brazila [[municipio]] kiu troviĝas en la administra unuo [[San-Paúlo]]. I K L	5.26%

Results of the editor study: We generate sentences that are highly reused by editors

---

## Conclusions

- Neural NLG approaches can work for underresourced languages with different properties (Arabic and Esperanto)
- Generated summaries are useful for article creation
- **Wikipedia's ArticlePlaceholder** is a good use case for NLG tasks.
- Engaging the community of readers / editors is the way to go when doing NLG for Wikipedia rather than automatic evaluation.

# Questions

---

- <https://tinyurl.com/y7opjsnl>  
Paper published at **ESWC 2018**
- <http://aclweb.org/anthology/N18-2101>  
Neural Network published at **NAACL 2018**



Lucie-Aimée Kaffee  
kaffee@soton.ac.uk



Hady Elsahar  
hadyelsahar@gmail.com



Pavlos Vougiouklis  
pv1e13@ecs.soton.ac.uk

## Mind the (Language) Gap: Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders

Lucie-Aimée Kaffee<sup>1†</sup>, Hady Elsahar<sup>2✉</sup>, Pavlos Vougiouklis<sup>1✉</sup>, Christophe Gravier<sup>2</sup>, Frédérique Laforest<sup>2</sup>, Jonathon Hare<sup>1</sup>, and Elena Simperl<sup>1</sup>

<sup>1</sup> School of Electronics and Computer Science, University of Southampton  
`{kaffee, pv1e13, jsh2, e.simperl}@ecs.soton.ac.uk`

<sup>2</sup> Laboratoire Hubert Curien, CNRS  
UJM-Saint-Étienne, Université de Lyon, France  
`{hady.elsahar, christophe.gravier, frederique.laforest}@univ-st-etienne.fr`

**Abstract.** While Wikipedia exists in 287 languages, its content is unevenly distributed among them. It is therefore of utmost social and cultural importance to focus efforts on languages whose speakers only have access to limited Wikipedia content. In this work, we investigate sup-