One World, One Wiki

C. Scott Ananian <cananian@wikimedia.org> [[User:cscott]] (Wikimedia Foundation)

Instead of today's many siloed wikis, separated by language and project, we aspire to re-establish a unified community of collaborators in the spirit of *ubuntu*. We will still respect language and cultural differences—there will still be English, German, Hebrew, Arabic, etc. Wikipedias; they will disagree at times—but instead of separate domains, we propose a single user experience with integrated navigation between projects and languages and the possibility of split screen views aligning related content. On a single page we can work on articles in different languages, or simultaneously edit textbook content and encyclopedia articles. Via machine translation we can facilitate conversations and collaborations spanning languages and projects, without forcing a single culture or perspective.

Language Converter

MediaWiki uses [[mw:LanguageConverter]] to automatically transliterate articles between closely related languages/dialects or script variants of a language/dialect. It is used on 11 wikis, and has been requested on about 35 more. Some examples of conversion pairs:

English (American/British) *LanguageConverter not used.* Spelling and usage differences exist between American English, British English, Indian English, and others.	[[**Elevator**]] An **elevator** is a type of vertical transportation that moves people or goods between floors (levels, decks) of a building, vessel, or other structure.	[[**Lift**]] A **lift** is a type of vertical transportation that moves people or goods between floors (levels, decks) of a building, vessel, or other structure.
Serbian (Latin/Cyrillic) *LanguageConverter in use.* Speakers are fully functionally digraphic, using both Cyrillic and Latin scripts. There are vocabulary differences between Ikavian, Ekavian, and Ijekavian dialects which are not currently converted.	[[**Lift**]] **Lift** je uređaj za transport ljudi ili tereta među spratovima zgrada ili radnih platformi. Uobičajno se kreće uz pomoć elektromotora koji ili pokreće užad za vuču i protivtežni mehanizam, ili pumpa hidrauličnu tečnost za podizanje cilindričnih klipova.	[[**Лифт**]] **Лифт** је уређај за транспорт људи или терета међу спратовима зграда или радних платформи. Уобичајно се креће уз помоћ електромотора који или покреће ужад за вучу и противтежни механизам, или пумпа хидрауличну течност за подизање цилиндричних клипова.
Chinese (Simplified/Traditional) *LanguageConverter in use.* Simplified used in mainland China, Singapore, and Malaysia. Traditional used in Taiwan, Hong Kong, Macau, and among Overseas Chinese. Most speakers monographic; few can fluently proofread text in both variants.	[[**电梯**]] **电梯**，亦称升降机、垂直电梯。在香港、新加坡和马来西亚俗称“转”（英语lift的译音），是一种垂直运送行人或货物的运输工具。	[[**電梯**]] **電梯**，亦稱升降機、垂直電梯。在香港、新加坡和馬來西亞俗稱「韓」（英語lift的譯音），是一種垂直運送行人或貨物的運輸工具。
Hindi/Urdu *LanguageConverter not used.* Urdu and Hindi are dialects of the Hindustani language, written in very different scripts: Arabic on the Pakistan side of the border, Devanagri on the India side. Currently separate small wikis; could combine efforts.	[[**उत्थापक**]] **उत्थापक**, उच्चालित्र अथवा एलिवेटर (lift या elevator) एक युक्ति है वस्तुओं एवं व्यक्तिओं को उर्ध्व दिशा में चढ़ाने-उतारने के काम आती है। प्रायः किसी बहुमंजिला ऊँचे भवन, जलपोत एवं अन्य संरचनाओं में उत्थापक लगा होता है जो गोलों को या सामान आदि को एक मंजिल से दूसरी मंजिल या एक स्तर से दूसरे स्तर पर लाता और ले जाता है। उत्थापक प्रायः विद्युत मोटर द्वारा चलते हैं।	[[**رافعہ**]] انتصابی نقل و حمل کی کل ہ جدید عمارتوں، جہ ازوں اور کانوں میں استعمال ہ ونے والی تمام کھلی اور بند ساختوں اور لگاتار چلنے والے ان پٹوں کو بھی رافع یا (انگریزی:Elevator) کہ ا جاتا ہے جو بھاری چیزوں کو ایک جگہ سے دوسری جگہ پہ نچاتے ہ یں ہ

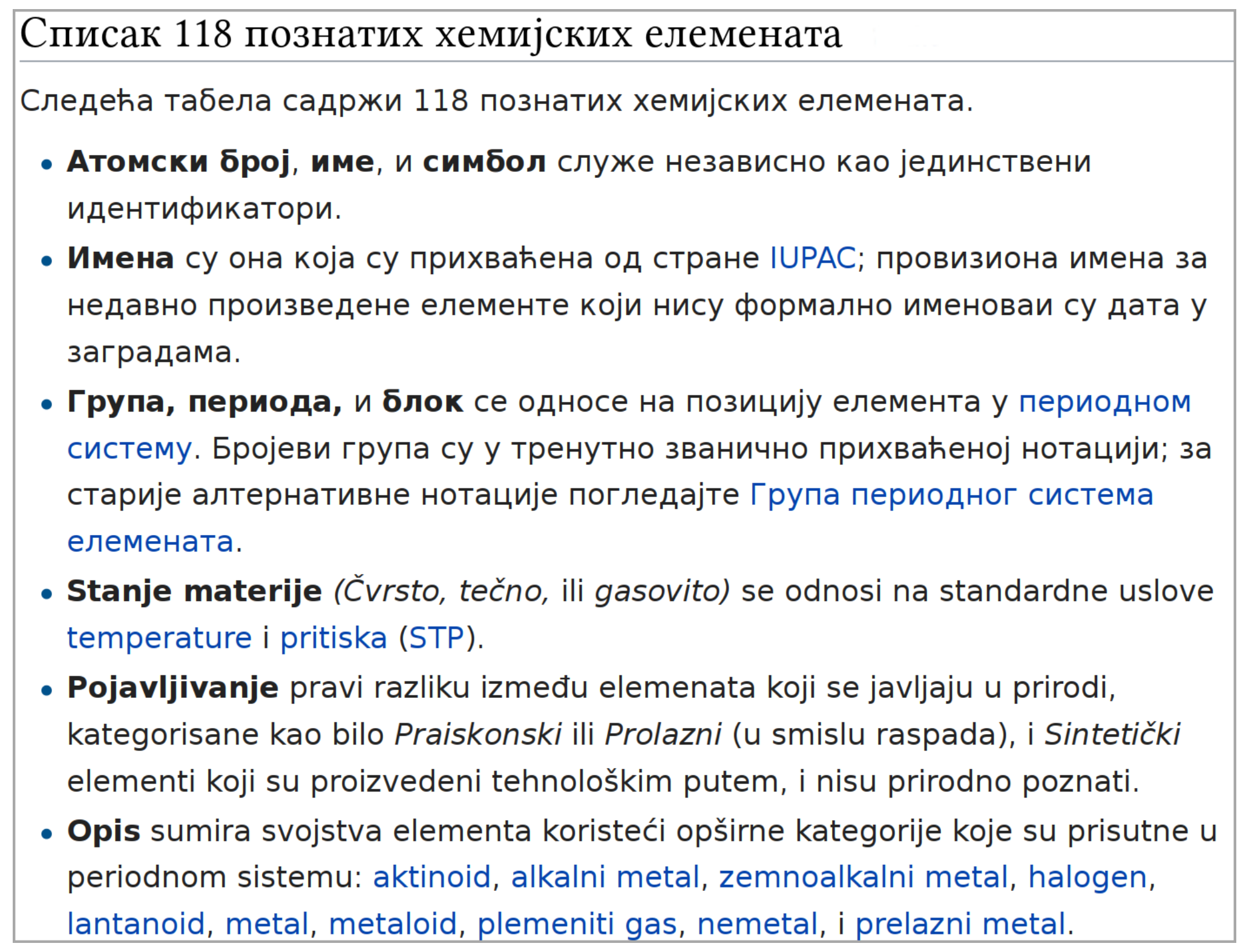
Native Variant Editing

LanguageConverter is oriented to readers: it converts the article text unidirectionally into readable text in a consistent variant. But as soon as a user begins to edit, they are confronted with the source text in a mix of variants, as illustrated by the intermingled Cyrillic and Latin scripts in the article from Serbian Wikipedia shown below. This mixture of scripts can be a huge barrier to editing in communities where individuals are typically only fluent in a single variant.

The Parsoid team has been experimenting with a new bidirectional implementation of LanguageConverter, based on Finite State Transducers (FSTs). These allow automatic annotation of wikitext such that it can be round-tripped to its original variant losslessly. With these annotations, an Wikimedian can edit an article in their preferred consistent variant. Unedited portions of the article will round-trip to their original variant, preventing dirty diffs.

On wikis where the community has chosen to author all articles in a single variant, all text can be losslessly saved as the chosen variant, regardless of which variant the editor used.

We can make editing easier on wikis using LanguageConverter!

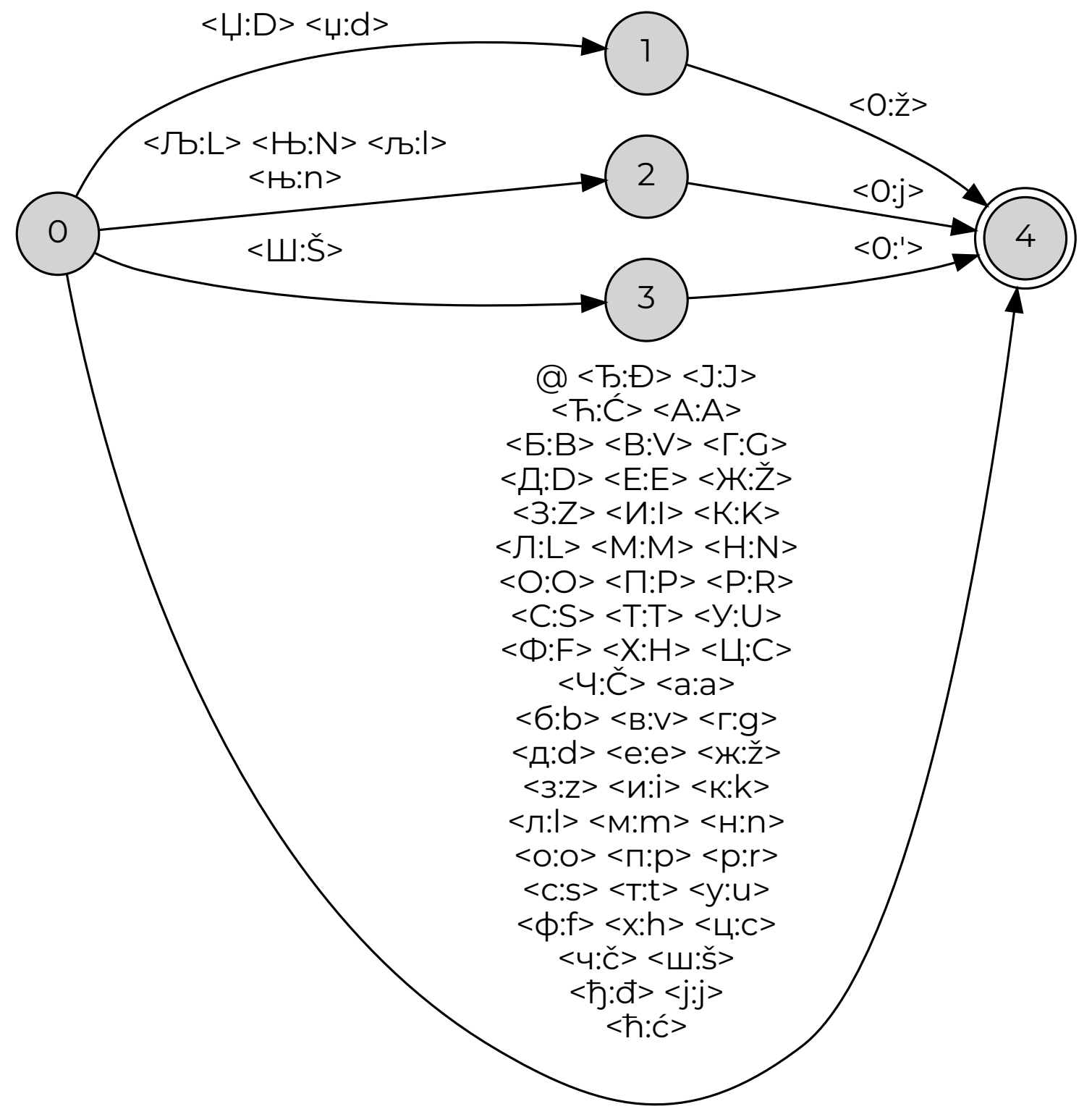


Списак 118 познатих хемијских елемената

Следећа табела садржи 118 познатих хемијских елемената.

- **Атомски број**, **име**, и **симбол** служе независно као јединствени идентификатори.
- **Имена** су она која су прихваћена од стране IUPAC; провизиона имена за недавно произведене елементе који нису формално именовани су дата у заградама.
- **Група, периода,** и **блок** се односе на позицију елемента у периодном систему. Бројеви група су у тренутно званично прихваћеној нотацији; за старије алтернативне нотације погледајте Група периодног система елемената.
- **Stanje materije** *(Čvrsto, tečno,* ili *gasovito)* se odnosi na standardne uslove temperature i pritiska (STP).
- **Pojavljivanje** pravi razliku između elemenata koji se javljaju u prirodi, kategorisane kao bilo *Praiskonski* ili *Prolazni* (u smislu raspada), i *Sintetički* elementi koji su proizvedeni tehnološkim putem, i nisu prirodno poznati.
- **Opis** sumira svojstva elementa koristeći opširne kategorije koje su prisutne u periodnom sistemu: aktinoid, alkalni metal, zemnoalkalni metal, halogen, lantanoid, metal, metaloid, plemeniti gas, nemetal, i prelazni metal.

Serbian Wikipedia article showing mixed Cyrillic/Latin script


0
1
2
3
4
<Џ:D> <џ:d>
<Љ:L> <Њ:N> <љ:l> <њ:n>
<Ш:Š>
<0:ž>
<0:j>
<0:'>
@ <Ђ:Đ> <Ј:J> <Ћ:Ć> <А:A> <Б:B> <В:V> <Г:G> <Д:D> <Е:E> <Ж:Ž> <З:Z> <И:I> <К:K> <Л:L> <М:M> <Н:N> <О:O> <П:P> <Р:R> <С:S> <Т:T> <У:U> <Ф:F> <Х:H> <Ц:C> <Ч:Č> <а:a> <б:b> <в:v> <г:g> <д:d> <е:e> <ж:ž> <з:z> <и:i> <к:k> <л:l> <м:m> <н:n> <о:o> <п:p> <р:r> <с:s> <т:t> <у:u> <ф:f> <х:h> <ц:c> <ч:č> <ш:š> <ђ:đ> <ј:j> <ћ:ć>

Partial FST for Serbian Cyrillic-to-Latin conversion

Translation Suggestion Tool

A translation suggestion tool **suggests an edit in one language whenever an edit is made to a parallel text in another language**. Correspondences are manually created via the Content Translation tool or "bandit learning", and machine translation is used to automatically create new (or prune old) correspondences.

Related work: `https://research.wikimedia.org/knowledge-gaps.html`

Zero-Shot Translation

"Zero-shot translation" machine translation models **allow training data from "big" wikis to improve the translation of "small" wikis**. Every contributed correspondence or translation further improves the ability of our tools to make additional articles from other languages available.

More information: `https://arxiv.org/pdf/1611.04558.pdf`

By bridging languages, our individual contributions can better fill knowledge gaps everywhere.