Language and Speech Technology
Technical Report Series

Report Number LST-16-01

Transcriptor

versie 0.01

Documentatie

Pepijn Hendriks, Martin Reynaert en Nicoline van der Sijs

February 1th, 2016 (date published) - February 1th, 2016 (last revision)

Language and Speech Technology PI Group
Centre for Language Studies
Radboud University Nijmegen
P.O. Box 9103
NL-6500 HD Nijmegen
The Netherlands
http://www.ru.nl/lst

Series editors:
Nelleke Oostdijk
Antal van den Bosch
David van Leeuwen

Inhoudsopgave

1 Inleiding **5**

2 Omzetten op twee manieren: snel of uitgebreid **9**

2.1 Algemeen 9

2.2 Accenten 10

3 Vlug Zoeken **11**

3.1 Transcriptie-Bezemer 11

3.2 Aardrijkskundige aanduidingen 12

3.3 Invoer (snelle methode) 13

3.4 Russisch (Cyrillisch schrift) 14

3.5 Engels (Latijns schrift) 14

3.6 Uitvoer (snelle methode) 15

3.6.1 Hoofdadvies 15

3.6.2 Overige schrijfwijzen 15

3.7 Beperkingen (snelle methode) 17

3.7.1 Variatie in het Engels 17

3.7.2 De ë, Nederlands 'jo' 18

3.7.3 Modern Russisch, modern Rusland 19

3.7.4 Geen andere talen 19

3.7.5 Overig 20

4 Uitgebreid Zoeken **21**

4.1 Bedoeling 21

4.2 Databanken 21

4.2.1 Persoonsnamen 22

4.2.2 Plaatsnamen 23

4.3 Werking: combinatie van fuzzy zoeken en op metadata gebaseerd zoeken 24

4.4 Input 24

4.5 Output 25

4.5.1 Algemeen 25

4.5.2 Uitgebreid zoeken over Russisch naar Nederlands 25

4.6 Traagheid Uitgebreid Zoeken 26

5 Technische uitvoering **29**

6 Verantwoording **31**

7 Contact **33**

Hoofdstuk 1

Inleiding

De Transcriptor zet namen, aardrijkskundige aanduidingen en woorden om van het Russisch naar het Nederlands.

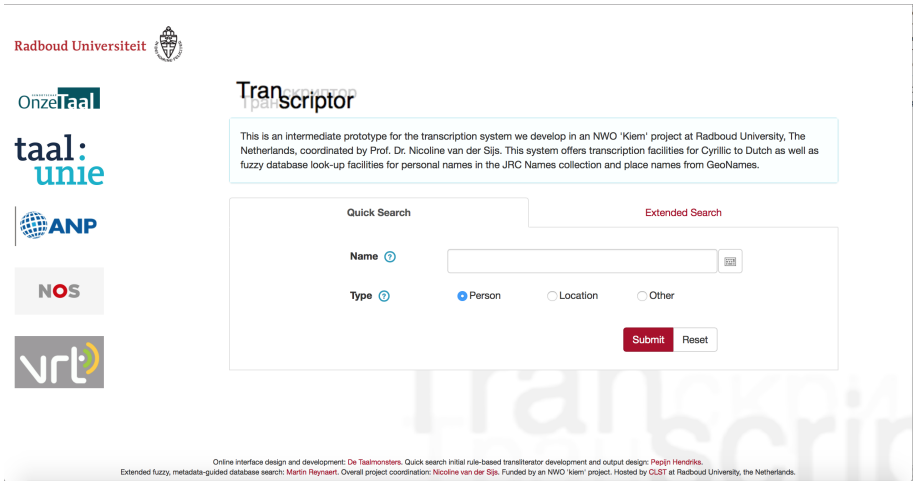
Als invoer kun je de originele Cyrillische schrijfwijze van zo'n naam, aanduiding of woord gebruiken – en dat werkt ook het best – maar de Transcriptor kan ook Engelstalige schrijfwijzen in het Latijnse alfabet omzetten naar de schrijfwijze die in het Nederlandse taalgebied gebruikelijk is.

Deze Nederlandse schrijfwijze die de Transcriptor genereert is in algemeen gebruik bij Nederlandse en Vlaamse media, in verzorgde Nederlandstalige non-fictie en fictie, bij de Nederlandstalige Wikipedia en bij slavisten en andere specialisten die in het Nederlands schrijven voor een algemeen publiek.

Een aantal grote Nederlandstalige spelers op dit gebied waren bij de ontwikkeling van de Transcriptor betrokken. De logo's van onze partners in dit project staan prominent op de opstartpagina van onze webapplicatie, zie Figuur 1.1.

Behalve de algemene Nederlandse schrijfwijze genereert de Transcriptor ook tal van andere schrijfwijzen. Het gaat dan om de schrijfwijzen die bijvoorbeeld gebruikelijk zijn in het Engels of het Duits, of die gebruikt worden door wetenschappers, in bibliotheken, op kaarten en plattegronden, langs de Russische wegen, in Russische identiteitsbewijzen en in het elektronische dataverkeer.

Doel van de Transcriptor is om gebruikers van het Nederlands de hand te reiken en te komen tot een consistente schrijfwijze van Russische namen, aardrijkskundige aanduidingen en woorden in Nederlandstalige teksten voor een algemeen publiek.



Figuur 1.1: Transcriptor opstartscherm

Deze tool is primair bedoeld voor journalisten, schrijvers, redacteuren, eindredacteuren en andere taalgebruikers die het Russisch (of het Cyrillische alfabet) niet machtig zijn, maar in de praktijk van alledag wel Russische persoons- en plaatsnamen tegenkomen – in het Russisch, maar nadrukkelijk ook in Engelstalige literatuur of persbureaukopij – en die namen en aanduidingen in een Nederlandse tekst willen gebruiken.

Tegelijkertijd is de Transcriptor ook een handig hulpmiddel voor taalgebruikers die het Russisch wél machtig zijn en snel en nauwkeurig een naam, aardrijkskundige aanduiding of woord volgens een bepaald systeem willen omzetten van het Cyrillische naar het Latijnse alfabet.

Tabel 1.1 geeft enkele voorbeelden van Cyrillische of Engelse input van de Transcriptor met de bijhorende Nederlandse output voor enkele persoons- en plaatsnamen.

Voorbeelden	
Invoer	Resultaat
Аркадий Гайдар	Arkadi Gajdar
Леонид Киселев	Leonid Kiseljov
Vladimir Putin	Vladimir Poetin
Alexander Pushkin	Aleksandr Poesjkin
Москва	Moskou
Челябинск	Tsjeljabinsk (Čeljabinsk)
Nizhny Novgorod	Nizjni Novgorod
Rashkino	Rasjkino

Tabel 1.1: Voorbeelden van Cyrillische en Engelse input en output van de Transcriptor voor persoons- en plaatsnamen

Hoofdstuk 2

Omzetten op twee manieren: snel of uitgebreid

2.1 Algemeen

Je kunt de Transcriptor op twee manieren aan het werk zetten: snel of uitgebreid. Bij de snelle methode wordt de invoer letter voor letter volgens vastgelegde regels omgezet. Bij de uitgebreide methode wordt aan de hand van een database met tal van bekende namen en aardrijkskundige aanduidingen de meest waarschijnlijke correspondentie bepaald.

Beide methoden hebben voor- en nadelen. De snelle methode heeft als belangrijkste voordeel dat ze – de naam zegt het al – snel is: ze genereert in een oogwenk een resultaat. Maar met name bij Engelstalige invoer loopt deze snelle methode tegen twee problemen aan. Ten eerste kan ten opzichte van het Russische origineel informatie verloren zijn gegaan. Daarnaast komt er in het Engels behoorlijk wat variatie voor in de schrijfwijze van sommige letters en klanken.

Bij het omzetten probeert de Transcriptor hiervoor zo goed en zo kwaad als het gaat te compenseren – wat met name bij veelvoorkomende voor- en achternamen ook aardig lukt – maar honderd procent zekerheid biedt de tool niet. De uitgebreide methode doet met name een groter beroep op je geduld: het kan wel enkele minuten duren voor je een resultaat krijgt. In de tussentijd wordt een grote database met persoonsnamen (JRC-Names) of aardrijkskundige aanduidingen (GeoNames) geraadpleegd en berekend welke bekende naam of aanduiding

het best past bij je invoer. Bij deze langzamere methode – het opzoeken vergt momenteel een tweetal minuten – zijn ook andere invoertalen dan Russisch of Engels mogelijk. Net als bij de snelle methode zijn ook de resultaten van de langzame methode echter lang niet altijd loepzuiver: met name de database JRC-Names – ontwikkeld door onderzoekers van de Europese Commissie ten behoeve van mediamonitoring – bevat nogal wat 'ruis'.

> Moet de Nederlandse schrijfwijze in je tekst absoluut goed zijn?
> Probeer van een Engelstalige schrijfwijze eerst het Russische origineel te achterhalen en gebruik dat als invoer.
> Accepteer het door het Vlugge Zoeken gegenereerde resultaat niet klakkeloos, maar zoek ermee verder in Google, Google Maps, Wikipedia of andere naslagwerken die je tot je beschikking hebt.
> Probeer echter eerst: Uitgebreid Zoeken.

2.2 Accenten

Als je een Russisch personage opzoekt in bv. Wikipedia krijg je in de Engelse versie ervan de Russische naam in het Cyrillisch met accenten terug. Deze accenten tonen waar bij het uitspreken van de naam de klemtoon ligt. Ze worden in het dagelijkse Russisch niet geschreven.

Ons systeem verwijdert deze accenten vooraleer het de gevraagde varianten in een andere taal gaat opzoeken.

Een uitzondering hierop is de trema op 'e'. Deze wordt in het dagelijkse Russisch meestal evenmin geschreven, maar heeft in feite wel een impact op de uitspraak van de naam. Waar de 'e' een trema zou krijgen in puristische orthografie – zoals in bv. leerboeken Russisch – verandert de uitspraak van 'e' of 'je' naar 'jo'. We hebben dit fenomeen zo goed mogelijk proberen op te vangen, in elk geval bij het 'Vlug Zoeken'. Hierover lees je in Sectie 3.7.2 meer.

Hoofdstuk 3

Vlug Zoeken

3.1 Transcriptie-Bezemer

Als basis voor de omzetting van het Cyrillisch naar het Nederlands wordt het systeem gehanteerd dat in 1973 werd opgesteld door de Amsterdamse hoogleraar Jan Willem Bezemer[1]. Deze 'transcriptie-Bezemer' komt voor de 33 letters van het Russische Cyrillische alfabet neer op de correspondenties zoals in Figuur 3.1.

In een aantal gevallen botst de omzetting volgens deze tabel – soms al dan niet met de kleine lettertjes – pontificaal met de spellingregels van het Nederlands.

Met name de combinatie van de Latijnse letters i en j is een grote spelbreker: zo levert 'Россия' ('Rusland' in het Russisch) volgens de tabel Rossija op, maar dit woord wordt in het Russisch zeker niet met een ij-klank uitgesproken. Een verder probleem is de combinatie j + j en een aantal gevallen van klinkerbotsing. Deze gevallen worden volgens een vast patroon opgelost[2]. Dit systeem wordt uitgelegd in Figuur 3.2.

[1] J.W. Bezemer, 'De Nederlandse transcriptie van Russische eigennamen', in: Internationale Spectator 27/2, 22 januari 1973, p. 37–39. Later ook overgenomen in zijn Een geschiedenis van Rusland. Van Rurik tot Breznjev, waarvan de eerste druk in 1988 verscheen bij uitgeverij G.A. van Oorschot in Amsterdam.

[2] Voor overige gevallen waarin strikt genomen sprake is van klinkerbotsing lijkt in het Nederlands een zekere gewenning te zijn opgetreden. Zo lijkt Michaïl te hebben plaatsgemaakt voor Michail. Vandaar dat in deze naam geen trema wordt geplaatst, en evenmin in een naam als Tsjaadajev. Als je streng in de leer bent, kun je in dergelijke gevallen natuurlijk gerust een trema plaatsen.

Аа	Бб	Вв	Гг	Дд	Ее	Ёё	Жж	Зз
a	b	v	g	d	e, je*	o, jo+	zj	z
Ии	Йй	Кк	Лл	Мм	Нн	Оо	Пп	Рр
i, ji†	j, -**	k	l	m	n	o	p	r
Сс	Тт	Уу	Фф	Хх	Цц	Чч	Шш	Щщ
s	t	oe	f	ch	ts	tsj	sj	sjtsj
Ъъ	Ыы	Ьь	Ээ	Юю	Яя			
-	y	-	e	joe	ja			

* *e* na medeklinker en и (*ië*);
je aan het begin van een woord, na ь en ъ,
en na andere klinkers dan и;

\+ *o* na ж (*zjo*), ч (*tsjo*), ш (*sjo*) en щ (*sjtsjo*); *jo* elders;
ook wanneer in het Russisch e in plaats van ё staat;

† *ji* na ь en ъ; *i* in overige gevallen

** geen transcriptie in de uitgangen -ый en -ий;
j in overige gevallen

Figuur 3.1: Overzicht van de lettercorrespondenties tussen de Russische en Nederlandse alfabetten

3.2 Aardrijkskundige aanduidingen

De schrijfwijze volgens bovengenoemde omzetting kan worden gebruikt voor alle namen van Russische personen en organisaties, aardrijkskundige aanduidingen en woorden.

Voor sommige aardrijkskundige namen bestaan echter geaccepteerde Nederlandse namen, zoals Moskou, Sint-Petersburg en Ladogameer. De Transcriptor geeft de Nederlandse namen weer die in het najaar van 2015 waren opgenomen in de lijst van de werkgroep Buitenlandse Aardrijkskundige Namen III (hierna: BAN III) van de Nederlandse Taalunie[3][4].

Voor overige Russische aardrijkskundige aanduidingen adviseert de werkgroep

[3]Zie `http://taaladvies.net/taal/aardrijkskundige_namen/land/RU`. De Transcriptor herkent ook enkele gangbare verkorte namen, zoals 'Бурятия' (Boerjatia) in plaats van 'Бурятская республика' (Boerjatskaja respoeblika) voor Boerjatië.

[4]De Transcriptor herkent daarnaast ook enkele veelvoorkomende Russische namen van plaatsen in Oekraïne. Het gaat om plaatsen die door de werkgroep bij dat land zijn opgenomen (zie `http://taaladvies.net/taal/aardrijkskundige_namen/land/UA`, waaronder ook plaatsen op de Krim. Dit autonome Oekraïense schiereiland wordt sinds maart 2014 door Rusland – maar niet door Nederland en België – beschouwd als Russisch grondgebied.

i + *j*	uitgang -ый, -ий	трудный, нижний ~~*trudnyj, nizjnij*~~ *trudny, nizjni*
	in *ija, ije, ijo*	Россия, Дмитриев, остриё ~~*Rossija, Dmitrijev, ostrijo*~~ *Rossia, Dmitriëv, ostrio*
	voor medeklinkers	Новороссийск, Брийск ~~*Novorossijsk, Brijsk*~~ *Novorossisk, Brisk*
j + *j*		Медвежье, Майя ~~*Medvezjje, Majja*~~ *Medvezje, Maja*
o + *oe*; *klinker* + *e*		Белоусов; аэропорт, Уэлен ~~*Belooesov; aeroport, Oeelen*~~ *Beloöesov*; *aëroport*, *Oeëlen*[2]

Figuur 3.2: Oplossing van klinkerbotsingen

BAN III de omzetting volgens een internationaal conversiesysteem dat bekendstaat als GOST-1983 (of GOST-83), en dat sinds 1987 wordt geadviseerd door de Verenigde Naties. Bij een aardrijkskundige aanduiding waarbij dit Taalunie-advies afwijkt van de omzetting volgens de transcriptie-Bezemer worden beide schrijfwijzen weergegeven.

3.3 Invoer (snelle methode)

Het is eigenlijk heel eenvoudig: vul het zoekveld in, geef aan of het om een naam, aardrijkskundige aanduiding of woord gaat en laat de Transcriptor zijn werk doen. Hou het eenvoudig: voer één naam per keer in en volg het reguliere hoofdlettergebruik.

Voor Russische invoer klik je op het toetsenbordicoontje rechts naast het invoerveld. Er verschijnt dan een virtueel Cyrillisch toetsenbord. Om je invoer af te sluiten klik je linksonder op het toetsenbordje op 'Accept'.

3.4 Russisch (Cyrillisch schrift)

De Transcriptor werkt het best als je een Russische naam, aardrijkskundige naam of woord in het originele Cyrillisch invoert. Dan zit alle informatie die nodig is voor een foutloze omzetting er nog in. Voer een naam in zijn grondvorm in: in de nominatief enkelvoud. De letter ë wordt in het Russisch vaak niet geschreven, en vervangen door een e. Voor de omzetting maakt dit nogal een verschil: jo versus e. Een aantal veelvoorkomende Russische voor- en achternamen met deze letter worden door de Transcriptor herkend. Zo worden 'Федор' en 'Киселев' herkend als 'Фёдор' en 'Киселёв' en correct omgezet als Fjodor en Kiseljov (niet: Fedor, Kiselev). Deze ingebouwde lijst is echter beperkt, vooral bij de achternamen. Voer voor een foutloze omzetting een naam, aardrijkskundige aanduiding of woord daarom in mét een eventuele ë (of probeer Uitgebreid Zoeken).

> De Transcriptor kan bij de snelle methode een ontbrekende ë in veel voornamen en enkele veelvoorkomende achternamen detecteren en invoegen.

3.5 Engels (Latijns schrift)

De Transcriptor doet ook zijn best om veelgebruikte Engelse schrijfwijzen om te zetten naar het Nederlands. Daarbij wordt zo veel mogelijk rekening gehouden met variatie die optreedt in verschillende Engelse omzetsystemen, zoals gebruikelijk in algemene teksten, maar ook in officiële Russische paspoorten.

Zo leveren Yury, Yuriy en Iurii alle drie Joeri (van 'Юрий') op, en wordt op basis van een ingebouwde lijst Evgeny en Ermolenko omgezet als Jevgeni en Jermolenko, terwijl Elvira geen j aan het begin krijgt en Elvira blijft.

Een ë in Russische namen, aardrijkskundige aanduidingen en woorden wordt in het Engels soms – en in Russische paspoorten eigenlijk altijd – als een e omgezet (en dus niet als o, yo of io). In de Engelse schrijfwijze van veel Russische voornamen kan de Transcriptor dit herkennen en de naam correct in het Nederlands weergeven: zo worden Petr, Artem en Fedor omgezet in Pjotr, Artjom en Fjodor. Hetzelfde geldt voor enkele veelvoorkomende Russische achternamen: Kiselev, Gorbachev en Seleznev bijvoorbeeld worden Kiseljov, Gorbatsjov en Seleznjov.

3.6 Uitvoer (snelle methode)

3.6.1 Hoofdadvies

Het belangrijkste advies dat de Transcriptor in de snelle methode geeft is de Nederlandse schrijfwijze die gebruikt kan worden in teksten voor een algemeen publiek.

Bij aardrijkskundige aanduidingen wordt ook een eventueel afwijkend Taalunie-advies gegeven, evenals een authentieke Nederlandse naam als die in de Taalunie-lijst voorkomt.

De Nederlandstalige Wikipedia gebruikt een schrijfwijze die op enkele punten licht afwijkt van het belangrijkste advies van de Transcriptor. Als dit zo is, wordt de Wikipedia-schrijfwijze apart aangegeven.

Met de links bij elk advies kan gemakkelijk met de geadviseerde schrijfwijze worden gezocht in Google, Google Maps en de Nederlandstalige Wikipedia.

3.6.2 Overige schrijfwijzen

Behalve het hoofdadvies genereert de Transcriptor bij invoer in het Cyrillisch een lange reeks andere schrijfwijzen.

> Bij invoer van een Engelse schrijfwijze in het Latijnse alfabet geeft de Transcriptor de meeste andere schrijfwijzen niet weer. Op basis van deze invoer kunnen ze simpelweg niet nauwkeurig genoeg worden gegenereerd.

- Engels, Duits
  - Gebruikelijke schrijfwijze in Engels- en Duitstalige teksten voor een algemeen publiek (niet uitputtend).

- Wetenschap
  - Gebruikelijke schrijfwijze onder slavisten en in specialistische wetenschappelijke publicaties.
  - ALA-LC: schrijfwijze volgens het ALA-LC-systeem van de American Library Association en de Amerikaanse Library of Congress. Veelgebruikt onder Engelstalige slavisten en in Engelstalige wetenschappelijke publicaties.
  - ALA-LC (versimpeld): schrijfwijze volgens het ALA-LC-systeem zonder diakritische tekens. Het weglaten van de diakritische tekens maakt de transliteratie leesbaarder, maar in sommige gevallen ambigu.
- Bibliotheken
  - Schrijfwijzen zoals gebruikt voor de catalogi van de Koninklijke Bibliotheek (Nederland), de Library of Congress (Verenigde Staten) en de British Library (Verenigd Koninkrijk).
    De aangeboden link laat onmiddellijk opzoeken van de transliteratie in de betreffende bibliotheek toe.
- Identiteitsdocumenten
  - Schrijfwijze zoals gebruikt in Russische paspoorten sinds 2013, op oudere paspoorten en op Russische rijbewijzen.
- Aardrijkskundige namen
  - VN, Nederlandse Taalunie: schrijfwijze aanbevolen door de Verenigde Naties en de Nederlandse Taalunie voor aardrijkskundige namen (bekend als GOST-1983 of GOST-83).
  - BGN/PCGN: Schrijfwijze aanbevolen voor officieel gebruik door de U.S. Board on Geographic Names (BGN; Verenigde Staten) en het

Permanent Committee on Geographical Names (PCGN; Verenigd Koninkrijk).

  - Russische wegwijzers: Schrijfwijze zoals voorgeschreven voor wegwijzers langs de Russische wegen. (In de praktijk lijkt deze schrijfwijze niet altijd strikt te worden gevolgd.)

- Internationale gegevensuitwisseling

  - ISO 9:1995: Schrijfwijze volgens de ISO-norm voor de uitwisseling van gegevens in het Cyrillisch, met name langs elektronische weg. Zeer strikte, altijd omkeerbare transliteratie.

  - De schrijfwijze volgens ISO 9:1995 is in Rusland en enkele andere ex-Sovjet-staten ingevoerd als systeem A uit de norm GOST 7.79-2000.

  - Behalve dit systeem A kent norm GOST 7.79-2000 ook een systeem B. Dit systeem is eveneens omkeerbaar.

De praktijk van identiteitsdocumenten: veel variatie
Op verzoek van de houder kan in een Russisch paspoort of op een Russisch rijbewijs ook de schrijfwijze uit een eerder paspoort of van een ander officieel document worden gehanteerd.
In paspoorten van voor 1997 werden Russische namen in het Latijnse schrift op een Frans aandoende manier geschreven. Deze schrijfwijze kent veel, niet altijd gedocumenteerde, uitzonderingen.

3.7 Beperkingen (snelle methode)

3.7.1 Variatie in het Engels

De Transcriptor probeert een Engelse schrijfwijze zo goed mogelijk om te zetten naar een Nederlandse (zie 3.5), ondanks de variatie binnen de verschillende sys-

temen die voor die taal in gebruik zijn. In een aantal gevallen wordt de variatie gewoon te groot.

De Russische varianten 'Наталия' (Natalia) en 'Наталья' (Natalja), bijvoorbeeld, kunnen in het Engels allebei geschreven worden als Natalia. De Transcriptor kan deze twee onmogelijk onderscheiden.

Hetzelfde geldt voor een Engelse lettercombinatie als vyu: het kan gaan om het Russische 'выу' (vyoe) of 'вью' (vjoe). Welke van de twee bedoeld wordt, kan de Transcriptor onmogelijk zeggen.

> Het compenseren voor variatie in de Engelse schrijfwijze van een Russische naam is bij de snelle methode uit de aard van het verschijnsel en de manier waarop wordt omgezet beperkt.
> Door deze beperking is een omzetting van een Engelse naar een Nederlandse schrijfwijze mogelijk niet helemaal correct, met name bij achternamen en aardrijkskundige aanduidingen

3.7.2 De ë, Nederlands 'jo'

Het probleem begint bij het Russisch – dat vaak een ë vervangt door een e – en zet zich voort in het Engels, waarin vaak niet o, yo of io wordt geschreven maar diezelfde e. In het Nederlands wordt traditioneel voor elke Russische ë een jo (soms o) geschreven (zie 3.1). De Transcriptor probeert bij invoer in zowel het Cyrillische als het Latijnse alfabet het ontbreken van de ë (of o, yo, io) te herkennen en daarvoor te compenseren. Dat lukt bij voornamen vrij aardig, maar blijft bij achternamen beperkt tot een handvol veel voorkomende. Voer voor een foutloze omzetting een naam, aardrijkskundige aanduiding of woord in het Cyrillisch daarom in mét een eventuele ë. Bij de Engelse schrijfwijze moet voor o, yo of io worden gekozen. We beseffen dat dit voor hen die het Russisch niet machtig zijn, vaak niet mogelijk zal zijn.

> De Transcriptor kan bij Vlug Zoeken een ontbrekende ë in veel voornamen en enkele veelvoorkomende achternamen detecteren en invoegen.

3.7.3 Modern Russisch, modern Rusland

De Transcriptor werkt voor persoonsnamen, aardrijkskundige aanduidingen en woorden in het Russisch en/of in Rusland. De spelling van het Russisch van vóór 1918, toen enkele letters werden geschrapt, wordt niet ondersteund.

3.7.4 Geen andere talen

Andere talen dan het Russisch die met het Cyrillische alfabet worden geschreven – zoals het Oekraïens, Wit-Russisch en Bulgaars, maar ook niet-Slavische talen als het Kazachs of het Karelisch – worden bij Vlug Zoeken niet ondersteund.

Wie met de Transcriptor een Engelse schrijfwijze van een Oekraïense of Wit-Russische naam of aardrijkskundige aanduiding omzet naar het Nederlands, kán geluk hebben. Maar garantie is er allerminst. De kwestie wordt gecompliceerd doordat in een land als Oekraïne en zeker in Wit-Rusland personen en plaatsen vaak onder twee namen bekendstaan.

Oekraïne (en Wit-Rusland)
Plaatsen in Oekraïne hebben behalve een officiële Oekraïense vaak ook een Russische naam. Soms is die Russische naam zelfs bekender. Denk aan de tweede stad van het land: die is bekender onder de Russische naam Charkov dan onder zijn Oekraïense naam Charkiv.
Voor personen geldt hetzelfde: de Oekraïense president Petro Porosjenko heet in het Russisch Pjotr Porosjenko. Omgekeerd heet de Russische nationale dichter Aleksandr Poejskin in het Oekraïens Oleksandr Poesjkin. De MH17 stortte in juli 2014 neer bij plaatsen die in het Oekraïens Hrabove en Rozsypne heten, en in het Russisch Grabovo en Rassypnoje.
De keuze voor de Oekraïense of de Russische naam van een plaats of persoon is geen spellingkeuze, maar een keuze tussen twee talen. In het Oekraïens en het Russisch wordt de keuze in de regel bepaald door de taal, en is deze keuze verder niet beladen. In Nederlandstalige teksten gaat het al snel om een politieke keuze: duid je een plaats aan volgens de officiële taal van het land, of in de taal die de meerderheid van de lokale bevolking spreekt en/of die van oudsher, uit de tijd van de Sovjet-Unie, het bekendst is?

In Nederland is de keuze die je in Oekraïne of Wit-Rusland moet maken te vergelijken met de keuze tussen Leeuwarden en Ljouwert. In België is ze te vergelijken met die tussen Schaarbeek en Schaerbeek, of tussen Boulevard Anspach en Anspachlaan.
In Wit-Rusland heeft het Russisch een hogere status dan in Oekraïne, maar afgezien daarvan speelt deze problematiek ook in dat land.
De Transcriptor kan bij deze keuze niet behulpzaam zijn.

3.7.5 Overig

In het Duits wordt een Cyrillische с tussen twee klinkers niet als s maar als ss omgezet: 'Исаев' wordt 'Issajew', niet 'Isajew'.

Tussen twee woorddelen lijkt deze regel niet op te gaan. Zoals in een plaatsnaam als 'Новосибирск', bestaande uit de elementen 'ново-' ('nieuw') en 'сибир-' ('Siberië'). Dit wordt in het Duits in de regel Nowosibirsk, niet Nowossibirsk. De Transcriptor houdt hiermee rekening bij invoer met de elementen sibir- ('Siberië'), nowo- ('Nieuw-'), leso- ('Bos-'), weliko- ('Groot-'), malo- ('Klein-'), en krasno- ('Rood-').

In de vroegere DDR werd de Cyrillische ж vaak niet met sch maar met sh omgezet, zodat een onderscheid kon worden gemaakt tussen bijvoorbeeld 'шарить' (scharit, 'op de tast zoeken') en 'жарить' (sharit, 'braden'). De Transcriptor maakt dit verschil niet.

Hoofdstuk 4

Uitgebreid Zoeken

4.1 Bedoeling

De bedoeling van het Uitgebreide Zoeken is dat iedereen een vreemde naam waarvan men geen idee heeft hoe die in feite gespeld zou moeten worden kan opzoeken. Het idee is dat dat men desnoods een naam die men alleen ergens gehoord heeft zou moeten kunnen intikken. Daarbij geeft men dan een zo goed mogelijke indicatie van wat men denkt de taal van oorsprong van de naam te zijn en verder in welke taal men de feitelijke schrijfwijze zou willen terugkrijgen. Wat men dan gehoord heeft en eerder fonetisch intikt en waarvan men denkt dat het oorspronkelijk een Russische naam was (zelfs al is het bv. in werkelijkheid een Oekraïense of een Poolse) en men dan graag in de Nederlandse schrijfwijze zou terugkrijgen, geeft men dan als volgt in: men tikt wat men gehoord heeft, men specificeert als 'tussentaal' het Russisch en als 'doeltaal' het Nederlands.

4.2 Databanken

Uitgebreid zoeken is gebaseerd op twee grote namendatabanken. Anders dan het op regelsets gebaseerde 'Vlugge' zoeken wordt bij 'Uitgebreid' dus uitgegaan van bestaande lijsten namen die verzameld werden in bijvoorbeeld de pers.

4.2.1 Persoonsnamen

JRC-Names bevat persoonsnamen uit het dagelijkse nieuws die verzameld worden door JRC[1] Steinberger et al. (2013). Hiertoe worden elke dag de media gecrawld en niet eerder geziene namen verzameld. Dit gebeurt systematisch voor alle landen die lid zijn van de Europese Unie, maar evengoed voor enkele niet-EU landen. Voor reeds eerder in de JRCnames opgenomen namen wordt bij opnieuw ontmoeten van de naam bijgehouden hoe vaak deze gezien werd. Het is ons momenteel niet duidelijk of deze frequentie de documentfrequentie dan wel de termfrequentie is, i.e. het aantal documenten waarin de naam gezien werd, dan wel hoe vaak de naam gezien werd los van het aantal documenten waarin de naam gezien werd.

De momenteel in de Transcriptor opgenomen namenlijst omvat 760.357 unieke naamsvarianten voor 338,634 unieke naamIDs voor in totaal 187 talen. Dit betekent niet dat elke naam in elke taal voorkomt. De naam van de beroemde Russische componist Tsjaikovski komt bijvoorbeeld wel voor in YY talen, maar komt niet voor in het Cyrillisch. Dit zou kunnen wijzen op verminderde populariteit in het huidige Rusland in vergelijking met andere landen, maar harde conclusies willen we hier niet aan verbinden. Een andere mogelijke verklaring zou evengoed kunnen zijn dat huidige Russische media een of andere populaire naam of omschrijving gebruiken om naar de man te verwijzen. Die populaire naam zou dusdanig van de officiele achternaam kunnen verschillen dat de matchingalgoritmes de link niet kunnen leggen.

De uit de JRC-Names ten behoeve van onze Transcriptor geëxtraheerde informatie per naam is de volgende: Voornaam, Familienaam, NaamID, frequentie, taalcode.

Op basis van de aan de naam toegekende ID wordt in JRCnames verband gelegd tussen de diverse talenvarianten van eenzelfde naam. De uitgebreide zoekfunctie van de Transcriptor maakt eveneens dankbaar gebruik van deze identifier om voor de variant van een naam in de ene taal, de variant in de andere taal op te halen.

Vreemde namen gegeven een specifieke taal kunnen best in verschillende varianten voorkomen. Het is precies de bedoeling van deze Transcriptor dat de verschillende partners in het project in de toekomst de door hen gehanteerde transcriptiesystemen en de daaruit voortvloeiende verschillende transcripties beter op elkaar afstemmen en stroomlijnen. Los daarvan komen in de pers wel eens

[1] `https://ec.europa.eu/jrc/en/language-technologies/jrc-names`

schrijffoutjes voor, ook in namen, en die leveren voor de JRC crawler dan weer een niet eerder geziene variant op. Als het goed gaat wordt het algoritme dat aan een naam de identifier toekent hier niet door in de war gebracht en krijgt de naam alsnog dezelfde identifier toegekend als de correcte spellingsvariant. Een enkele keer gaat dit mis en wordt als het ware een nieuwe naam met nieuwe ID gecreëerd. Voor Tsjaikovski vinden we 5 varianten met naamID 'sense726213' en 6 varianten met NaamID 'sense938244'. Voor echt hedendaagse mediafiguren kunnen deze aantallen varianten oplopen tot in de honderden.

4.2.2 Plaatsnamen

Voor plaatsnamen maken we in de huidige Transcriptor gebruik van aan het project GeoNames[2] ontleende lijsten met plaatsnamen. Zo is er de lijst 'cities1000' die in principe de namen bevat van alle steden en dorpen die meer dan 1000 inwoners tellen, waar ook ter wereld. Naast deze lijst is een aparte lijst 'alternatenames' beschikbaar waarin de bekende varianten van de plaatsen in andere talen opgenomen zijn. Bij deze lijst komt bij elke variant ook de ISO-code van de taal van de variant voor. Deze lijsten zijn gelinkt via een numerieke ID van elke naam.

De lijst cities1000 geeft 144.861 unieke NaamIDs. De lijst alternate names heeft echter 4.792.603 unieke IDs en omvat bijgevolg veel meer informatie dan enkel de grotere dorpen en stedennamen. Als extra metadata geeft het overigens niet alleen de taalcodes voor de varianten maar ook bv. postcodes en de Wikipedia URL, zo die bestaat. Aan unieke metadatatags, vooral talennamen, tellen we er 438.

Voor gebruik in de Transcriptor extraheren we uit deze lijst op basis van de IDs de volgende informatie: ID, plaatsnaam, inwonersaantal en taalcode. Bij plaatsnamen gebruiken we dus in feite de inwonersaantallen op dezelfde manier als de frequentie bij de persoonsnamen.

[2] `http://www.geonames.org/`

4.3 Werking: combinatie van fuzzy zoeken en op metadata gebaseerd zoeken

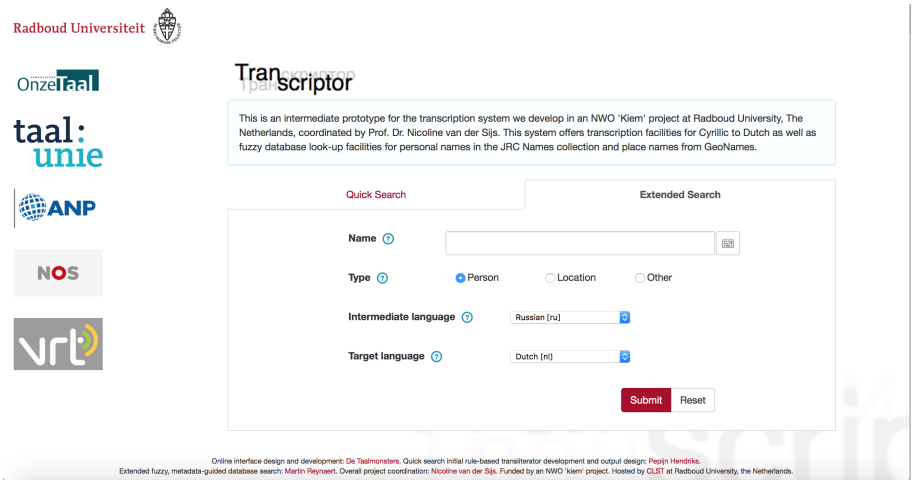
De NaamIDs en de frequenties/inwondersaantallen zijn aanvullende informatie bij de namen, meer technisch: metadata genoemd.

Enerzijds hebben we dus vooral bij de persoonsnamen een onbekend aantal naamsvarianten per taal voor een onbekend aantal talen (niet elke naam is immers bekend in elke taalvariant). Anderzijds biedt de JRCnames ons geen informatie over welke variant van een naam nu in werkelijkheid de correcte variant zou zijn. Daarom gebruiken we de frequentie als substituut, we nemen aan dat de variant die het vaakst voorkomt de correcte vorm van de naam is. Bij het Russisch geeft dat een extra complicatie. Het Russisch kent naamvallen en zo is het goed mogelijk dat bijvoorbeeld de genitiefvorm (bv. Александра Солженицына, i.e. 'van Alexander Solzjenitsyn') vaker voorkomt dan de nominatief, de basisvorm. Daarom geeft Uitgebreid Zoeken de drie meest frequente vormen van een naam terug, hoe ze ook verbogen zijn.

4.4 Input

Figuur 4.1 toont de interface voor het Uitgebreide Zoeken in de Transcriptor. Hierbij zijn in vergelijking met het Vlugge Zoeken opstartscherm twee extra velden opgenomen. Default zijn deze ingesteld om via het Russisch te zoeken naar de Nederlandse varianten van de naam die u wilt opzoeken.

In tegenstelling tot het Vlugge Zoeken biedt het uitgebreide zoeken echter de mogelijkheid met tientallen andere taalcombinaties te werken. U kunt net zo goed met als tussentaal Chinees opzoeken hoe men 'Mao Tsedong' schrijft in het Spaans. Het originele Chinees wordt als bonus eveneens teruggegeven.



Figuur 4.1: Transcriptor: opstartscherm Uitgebreid Zoeken

4.5 Output

4.5.1 Algemeen

Anders dan de output van het Vlugge Zoeken dat in HTML per categorie transliteratie zijn output levert, geeft Uitgebreid Zoeken momenteel zijn output als platte tekst terug.

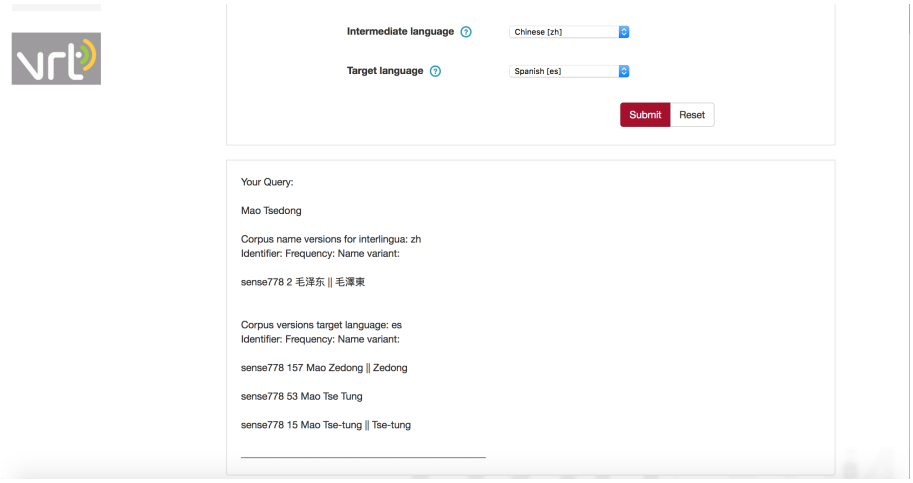
Indien gevonden, worden tot 3 varianten voor elke inputvraag teruggegeven. Deze zijn van elkaar afgescheiden door een verticale streep.

Figuur 4.2 toont de output (en de onzekerheid wat betreft de schrijwijze in het Spaans) van bovenbeschreven voorbeeld betreffende query 'Mao'.

4.5.2 Uitgebreid zoeken over Russisch naar Nederlands

De taalkeuzes bij Uitgebreid Zoeken staan default ingesteld voor de onmiddellijke doelgroep van de huidige Transcriptor. Zo staat de tussentaal ingesteld op Russisch, de doeltaal op het Nederlands.

Zoek je op een bekende Russische voornaam, vroeger ook vaak in het Nederlands zoals nog steeds in het Duits met 'W' gespeld, bv. Wladimir, dan zal het systeem



Figuur 4.2: Transcriptor: output Uitgebreid Zoeken, query 'Mao Tsedong'

de drie meest voorkomende mediaprominenten met die voornaam teruggeven. Momenteel zijn dat dan de Vladimirs: Poetin, ... en ...

Zoek je daarentegen op een bekende achternaam, dan is het best mogelijk dat samen met de persoon die je wellicht in gedachten had, bv. zijn vrouw of andere familieleden of naamsverwanten eveneens teruggeven worden. (vb. Solzhenitsyn).

4.6 Traagheid Uitgebreid Zoeken

De databanken voor Uitgebreid Zoeken vereisen wanneer operationeel elk 3 tot 4 gigabytes in het werkgeheugen van de server waarop de Transcriptor draait. Beide databanken worden daarom apart en nooit gelijktijdig geladen. Om dezelfde reden worden ze niet online gehouden, onze huidige infrastructuur laat ons niet toe deze hoeveelheid werkgeheugen permanent beschikbaar te houden terwijl in verhouding onze applicatie wellicht slechts sporadisch geraadpleegd zal worden.

Bij raadpleging van Uitgebreid Zoeken wordt dus alles van schijf ingelezen en voorbereid om de fuzzy zoekvraag te ontvangen. Dit vraagt tijd. In de praktijk vraagt dit dus voor JRC-Names ongeveer 2 minuten, voor GeoNames volgens onze testen net iets langer.

We hopen deze wachttijden in de toekomst in een vervolgproject drastisch te

kunnen inkorten.

Hoofdstuk 5

Technische uitvoering

- Technische realisatie van deze webapplicatie is te danken aan CLAM van Gompel (2012). Deze 'Computational Linguistic Application Mediator' zorgt er in feite voor dat deze webapplicatie evengoed als webservice aangesproken kan worden van Gompel and Reynaert (2014). Bij deze applicatie werd voor het Vlugge Zoeken gebruik gemaakt van de CLAM voorziening 'Actions'. Het Uitgebreide Zoeken steunt op de CLAM voorziening 'Projects'.

- TICCL back-end Uitgebreid Zoeken. Het fuzzy zoeken in de Uitgebreide methode is gebaseerd op een aantal modules van het systeem TICCL (Text-Induced Corpus Clean-up) Reynaert (2010). Hierbij is de zogenaamde focuswoordaanpak gehanteerd: gegeven de ingevoerde query wordt exhaustief gezocht naar alle varianten in de databank die binnen een editeerafstand van twee letters vallen. De flinke hoeveelheid ruis in de databanken kan op deze manier zelfs assisterend werken. Waar enkel correcte vormen vaak de editeerlimiet tussen fonetisch ingevoerde queries zouden te buiten gaan, vallen minder correcte vormen soms binnen het bereik en leiden alsnog naar de meer gebruikelijke vorm.

Hoofdstuk 6

Verantwoording

De Transcriptor werd in 2015 ontwikkeld door dr. Pepijn Hendriks (slavist) en dr. Martin Reynaert (computationeel taalkundige) in het kader van het KIEM-project Transcriptie App voor Cyrillisch schrift, onder leiding van prof.dr. Nicoline van der Sijs. Het project werd uitgevoerd aan het Centre for Language and Speech Technology (CLST) van de Radboud Universiteit Nijmegen en gefinancierd door het NWO. De interface is ontworpen door Matje van de Camp van De Taalmonsters, met extra financiering vanwege CLST.

Externe partners bij dit project waren de Nederlandse Taalunie, het Genootschap Onze Taal, de NOS, het ANP en de VRT. Bij de totstandkoming van de Transcriptor is dankbaar gebruikgemaakt van hun inbreng, advies en commentaar. Uiteraard kunnen deze organisaties op taalkundige, journalistieke of andere gronden steeds hun eigen afwegingen maken en zijn ze niet gebonden aan de door de Transcriptor aanbevolen schrijfwijzen.

Hoofdstuk 7

Contact

Voor vragen over het KIEM-project Transcriptie App voor Cyrillisch schrift kun je contact opnemen met Nicoline van der Sijs: post@nicolinevdsijs.nl.

Voor vragen over de aanbevolen schrijfwijzen en de omzetting volgens de snelle methode kun je contact opnemen met Pepijn Hendriks: pepijn@pepicek.nl.

Voor vragen over de techniek en over de omzetting volgens de uitgebreide methode kun je contact opnemen met Martin Reynaert: reynaert@uvt.nl.

Voor vragen over de interface kun je contact opnemen met Matje van de Camp van De Taalmonsters: matje@taalmonsters.nl.

Bibliografie

Reynaert, M. (2010). Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14:173–187. 10.1007/s10032-010-0133-5.

Steinberger, R., Pouliquen, B., Kabadjov, M. A., and der Goot, E. V. (2013). Jrc-names: A freely available, highly multilingual named entity resource. *CoRR*, abs/1309.6162.

van Gompel, M. (2012). CLAM: Computational Linguistics Application Mediator. Documentation. Technical report, Tilburg Centre for Cognition and Communication, Tilburg University and Radboud Centre for Language Studies, Radboud University Nijmegen.

van Gompel, M. and Reynaert, M. (2014). CLAM: Quickly deploy NLP command-line tools on the web. In *Proceedings of COLING 2014: System Demonstrations*, pages 71–75, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.