Intermediate representation in rule-based machine translation for the Uralic languages [*]

Francis M. Tyers,
HSL-fakultehta
UiT Norgga árktalaš universitehta
francis.tyers@uit.no

Tommi A. Pirinen
ADAPT Centre
School of Computing,
Dublin City University
tommi.pirinen@computing.dcu.ie

Abstract

This paper presents some of the major obstacles and challenges in creating machine translation systems between Uralic languages where the intermediate representation is based on morphology and syntax. The Uralic languages are very alike in many ways: similar case inventories, word order and non-finite clause forms. However current rule-based grammatical resources take many different approaches to encoding this information. These approaches are sometimes based on legacy or traditional grammatical description, important for making the tools comfortable for linguists, but sometimes based on arbitrary and incompatible decisions. This paper presents an overview of some of the issues in working with existing tools and representations and provides some guidelines and suggestions to facilitate future work.

1 Introduction

Creating *rule-based machine translation* (RBMT) systems is a process where one creates a mapping between units of source language and target language. The units can be different depending on the approach to the problem, i.e., on scale of translating word-forms to word-forms to translating via an intermediate abstract universal language, or an *interlingua*. In this article we study the approach of using just morphological analysis with the Uralic languages. The problem of such a system is that, even when morphologies of the closely related Uralic languages are expected to match, there are often engineering issues that make the work more tedious and cumbersome than necessary. Minimising the amount of simple engineering work is vital for making rule-based machine attractive to linguists and programmers alike.

The rest of the article is structured as follows: first we describe the backgrounds of the problem in 2, then we introduce the resources we are going to use in 3, we suggest some common best practices in 6, in 7 we briefly describe universal parts-of-speech and morphological features, and finally in 8 we provide some short concluding remarks.

[*] This article has a free/open publication licence; this particular version is Tommi A Pirinen's draft version. This version uses my own documentclass instead of official publication's, if any. This version is optimised for screen reading. This version uses only free fonts if possible. This work is licensed under a Creative Commons Attribution–NoDerivatives 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by-nd/4.0/. Original publication in proceedings of second IWCLUL held in Szeged 2016

James	*ja*	*Mary*	
+N+Prop+Sem/Mal+Sg+Nom	+CC	+N+Prop+Sem/Fem+Sg+Nom	
leaba	*gárdimis*	.	
+V+IV+Ind+Prs+Du3	+N+Sg+Loc	+CLB	
Джеймс	*марто*	*Марит*	
+N+Prop+Sem/Mal+Sg+Nom+Indef	марто+Po+COM	+N+Prop+Sem/Fem+Pl+Nom+Indef	
садпиресэть	.		
+N+SP+Ine+Indef+Der/Pr+V+Ind+Prs+ScPl3	+CLB		
James	*ja*	*Mary*	*ovat*
N Prop Nom Sg	Part	N Prop Nom Sg	V Prs Act Pl3
puutarhassa	.		
N Ine Sg	Punct		
James	*ja*	*Mary*	*on*
+H+sg+nom	+J	+H+sg+nom	+V+indic+pres+ps3+pl+ps+af
aias	.		
+S+sg+in	.		
James	*és*	*Mary*	a
/NOUN	/CONJ	/NOUN	/ART
kértben	*vannak*	.	
/ADJ<CAS<INE»	/VERB<PLUR>	/PUNCT	

Table 1: Translations of the sentence 'James and Mary are in the garden.' in several Uralic languages (North Sámi, Erzya, Finnish, Estonian, Hungarian) with the tag strings used in their morphological analysers. There are examples of real morphosyntactic differences (compare the third-person dual in North Sámi with the third-person plural in other languages) and arbitrary tag differences (compare the tag that the word for *and* receives in the different languages).

2 Background

RBMT is a popular way of developing high-quality machine translations between related languages [1]. The building of an RBMT system rapidly for related languages is possible, as has been done with, e.g. Dutch and Afrikaans [2]. A wide-coverage machine translation requires wide-coverage lexical resources for the languages. Developing an analyser to a stage where it is usable by multiple applications, including RBMT, can take years, so it is often a good idea to use readily available resources instead of re-writing a new analyser from the scratch. However, the majority of existing analysers are made with language-dependent annotation systems, which unnecessarily complicate the description of machine translation. It should be clear, that if two related languages use the same morphological and syntactic structures to describe a phenomenon, a rule mapping between the two should be entirely trivial. This is not the case when taking most off-the-shelf analysers for contemporary Uralic morphologies. Table 1 shows an example of the morphological annotation of five Uralic languages for a simple five-word sentence.

2.1 Intermediate representations

In machine translation, an intermediate representation is an abstraction away from the surface forms of the language. Figure 1 shows the Vauquois triangle, a common illustration of different levels of intermediate representation.

At the bottom of the triangle, there is no intermediate representation and translation is performed on a word-for-word basis. At the apex of the triangle is interlingual translation, where the source language is first mapped to a language-independent semantic representation, and this representation is then used to generate the target language.

In the middle is (morpho-)syntactic transfer. Here the source language is analysed to a language-dependent intermediate representation (usually based on a combination

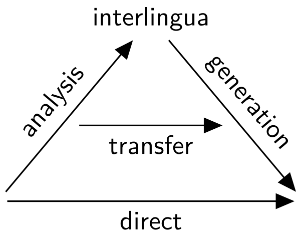

Figure 1: The Vauquois triangle which illustrates the amount of transfer needed for different levels of intermediate representation.

of syntactic structure and morphosyntactic features) and then transfer rules are applied to convert the source language intermediate representation to one compatible with the target-language generation component.

3 Resources

In this paper we make use of five sets of linguistic data for five different Uralic languages: Finnish, North Sámi, Erzya, Estonian and Hungarian. We take the North Sámi and Erzya data from the Giellatekno language technology repository.[1] The North Sámi data has primarily been developed by the Divvun and Giellatekno groups at UiT Norgga árktalaš universitehta and the Erzya data has been developed by Jack Rueter at Helsingin yliopisto [3]. For the Estonian data, we use the *plamk* analyser[2] written by Jaak Pruulmann-Vengerfeldt, for Finnish, *omorfi* [4][3] and for Hungarian, hunmorph [5].[4]

4 Strategies

There a different ways to fix systematic mismatches. We evaluate the followings:

4.1 Relabelling

An obvious approach to getting around the problem of divergent tagsets is to simply perform relabelling. This is where you replace the canonical tags in one language with their equivalents in the other language, or with a common equivalent in both languages.

$$+\text{CC} \rightarrow \text{<cnjcoo>} \leftarrow +\text{J}+\text{Coord}$$

However, this solution has its disadvantages. Even though +J and +CC both are used for conjuctions, the *plamk* tag is also used with subordinating and other conjunctions, while the Giellatekno tag excludes those. Relabelling +J+Coord to +CC and any other +J to +CS might work on the analyser, but will not work in a disambiguation rule saying “select the noun reading if the word to the right is tagged +J”, here we need to relabel +J to (+CS or +CC). In the opposite direction, +CS would need to be relabelled to (+J

[1] `http://giellatekno.uit.no`

[2] `https://github.com/jjpp/plamk`

[3] `https://github.com/flammie/omorfi`

[4] `http://mokk.bme.hu/resources/hunmorph/`

but not +Coord). The distinction between these may be irrelevant for the translation process (in all cases, *ja* in North Sámi will be translated to *ja* in Estonian), but for the intervening grammatical tools, it may be vital to make (or not) the distinction.

4.2 Interlingua

Another potential solution is to use a semantic interlingua (see description in section 2.1). This is the approach adopted by the machine translation system based on Grammatical Framework [6].[5] In this framework there is no direct transfer of morphological features.

5 Specific linguistic issues

There are a number of linguistic issues in RBMT. We cover the following in detail:

5.1 Copula

There are two main copula constructions in the Uralic languages, the first functions more or less like in the Germanic languages. The copula is a normal verb that agrees with the subject. The second copula construction works like in the Turkic languages. In languages with the Turkic-style copula, it does not typically surface in the third-person singular present tense. In our examples, North Sámi, Finnish and Estonian are of the Germanic type, while Hungarian and Erzya are of the Turkic type.

	‘She is a student.’	‘She was a student.’
North Sámi	Son lea studeanta.	Son lei studeanta.
Erzya	Сон студент.	Сон студентель.
Finnish	Hän on opiskelija.	Hän oli opiskelija.
Estonian	Ta on üliõpilane.	Ta oli üliõpilane.
Hungarian	Ő hallgató.	Ő hallgató volt.

In North Sámi, Finnish and Estonian, the treatment of *lea, on* is similar. It is a verb which inflects and agrees like other verbs.

There are divergences when we look at the Erzya and Hungarian examples. Although they have the same structure, zero copula in the present tense and surfaced copula in the past tense. The morphological analyser for Erzya treats the copula as a derivation:

студент+N+Sg+Nom+Indef+Der/Pr+V+Ind+Prs+ScSg3

Where in Hungarian it is simply omitted in the present (if it surfaced it would be *van*), and in the past it is considered a verb form.

5.2 Non-finite verb forms

Non-finite verb forms are infinitives and participles on the on hand and derivations on the another. There are a different number of them between languages and their tasks vary from being syntactic arguments of constructions to derived words, and a wide range of analyses are used to accommodate that. There are some differences in the table 2

[5] `http://grammaticalframework.org`

Language	Sentence	Non-finite tag
	‘I see the man who is running’	
North Sámi	Oidnen dievddu viehkame	Actio+Ess
Erzya	Неян цёранть, конась чийни.	Der/Ы+ActPrcShort+A
Finnish	Näen miehen juoksemassa.	InfMA+Ine
Estonian	Näen meest, kes jookseb.	—
Hungarian	Látom a futó embert.	/VERB[IMPERF_PART]/ADJ
	‘While running I saw the man’	
North Sámi	Oidnen dievddu viegadettiinan.	Ger+Px1Sg
Erzya	Неян чийниця цёранть.	Der/Ыця+ActDemPrc+A
Finnish	Näin miehen juostessani.	InfE+Ine+PxSg1
Estonian	Jooksmise ajal nägin ma meest.	Der/mine+Gen
Hungarian	Futás közben láttam az embert.	/VERB[GERUND]/NOUN
	‘I see the running man.’	
North Sámi	Oainnán viehkki dievddu.	PrsPrc
Erzya	Чийнемась седень кецявты.	Der/ОмA+Nom
Finnish	Näen juoksevan miehen.	PrsPrc
Estonian	Näen jooksvat meest.	Der/v+A+Nom
Hungarian	Látom a futó embert.	/VERB[IMPERF_PART]/ADJ
	‘Running is fun.’	
North Sámi	Viehkan lea suohtas.	Actio+Nom
Erzya	Мелезэнь тукшны чийнемась.	Der/ОмA+Nom
Finnish	Juokseminen on kivaa.	Der/minen+Nom
Estonian	Jooksmine on lahe.	Der/mine+Nom
Hungarian	A futás jó dolog.	/VERB[GERUND]/NOUN
	‘I like running.’	
North Sámi	Liikon viehkat.	Inf
Erzya	Чийнемстэ неия цёранть.	Inf+Ela
Finnish	Pidän juoksemisesta.	Der/minen+Ela
Estonian	Mulle meeldib joosta.	Inf
Hungarian	Szeretem futni.	/VERB<INF>

Table 2: Examples of the use and tagging of non-finite verb forms in the languages in our sample. It is not to be expected that the tags are completely equivalent, but for example, given the similarity in structure, should there be a difference in annotation between Finnish PrsPrc and Estonian Der/v+A?

5.3 Derivation, compounding and lexicalisation

A classical problem in computational morphologies lies in question of lexicalisation and productivity of certain processes; is a morphologically created word-form a new word or a form of a, possibly distant root. Morphologies take widely different and opposing approaches to this ranging from lexicalise-everything to collect-everything. See examples below:

	‘to drink’	‘a drink’	‘drinker’	‘brewery’
North Sámi	juhkat	juhkamuš	—	vuolla·buvttadeaddji
Erzya	симемс	симема-пель	симиця	пиянь завод
Finnish	juoda	juo-ma	juo|ja	olut·tehdas
Estonian	jooma	joo|gi	joo|	õlle·tehas
Hungarian	iszik	ital	iv|ó	sör·főzde

The symbols ‘·’, ‘-’ and ‘|’ stand for compounding, inflection and derivation, respectively.

5.4 Pronouns and determiners

The distinction between pronoun and determiner is not widely made in traditional grammars of most Uralic languages. Words which may be considered both pronouns and determiners are lumped into a single morphosyntactic class (usually pronoun). Consider the following examples involving the word ‘this’

	‘I see this house.’	‘I see this.’
North Sámi	Oainnán dán viesu.	Oainnán dán.
Erzya	Неян те$_{\text{det}}$ кудонть.	Неян тень$_{\text{pron}}$.
Finnish	Mä näen tämän$_{\text{pron}}$ talon.	Mä näen tämän$_{\text{pron}}$.
Estonian	Ma näen selle$_{\text{pron}}$ maja.	Ma näen selle$_{\text{pron}}$.
Hungarian	Nézem ezt$_{\text{det/noun}}$ a$_{\text{art}}$ házat.	Nézem azt$_{\text{det/noun}}$

In traditional grammars of North Sámi, Finnish and Estonian both the pronominal and the modifier analyses of ‘this’ are classified as pronouns. In Hungarian and Erzya, a distinction is made, with Hungarian making a pronoun/determiner distinction and Erzya making a distinction between quantifier (determiner) and nominalised quantifier.

If we consider a standard definition of *pronoun* to be ‘that which stands in place (pro-) of a noun phrase (-noun)’ then we can see that in the above, only the tools for Erzya follow this. The other languages leave the distinction to tools later in the pipeline.

5.5 Non-inflecting words

All languages in the Uralic family have a wide variety of non-inflecting word forms. Depending on the grammatical tradition followed by the language resource these may be simply lumped into a single class, or they may have extensive syntactic or semantic subcategorisation. Table 3 gives a number of examples of non-inflecting words and the equivalent morphological analyses they receive in each of the languages we are studying. To a machine translation practitioner, these distinctions are largely superfluous, *ja* in North Sámi will be translated as *ja* in Finnish and *ja* in Estonian. However, the distinctions may be vital for the intervening disambiguation tools, and as such need to be taken into account.

	North Sámi	**Erzya**	**Finnish**	**Estonian**	**Hungarian**
and	ja+CC	марто+Po+COM	ja Part	ja+J	és /CONJ
very	hui+Adv	пек+Adv+AdA	tosi Part	väga+Adv	nagyon /ADV
under	vuolde+Po	алов+Po+Lat	alle Part	alla+K	alatt /POSTP
now	dál+Adv	ней+Adv+Temp	nyt Part	praegu+Adv	most /ADV
hello	bures+Interj	шумбрачи+Interj+Formulaic	moi Part	tere+I	szia /UTT-INT

Table 3: Some examples of non-inflecting words with divergent morphological and syntactic annotation. In terms of morphology, the transfer of these tags may be a simple one-to-one substitution. However the syntactic environments may vary substantially.

6 Guidelines

6.1 Separation of lexicon and morphotactics

One of the main components of any rule-based system for morphologically-complex languages is a lexicon consisting of stems and inflectional/derivation categories. In some cases, such as for Finnish, these are partly provided by a state institution, such as a language board. In other cases they are the product of many years of work.

Although categorising stems for inclusion in a morphological lexicon (many contain over 100,000 entries) can take a substantial amount of work, even if done semi-automatically, implementing the morphotactics (that is, the rules covering inflection, derivation and compounding) may take substantially less time.

6.2 Maximise parallelism

In line with the Universal Dependencies project (see 7), we propose the adoption of a principle of maximum parallelism. In short “things that are the same should be tagged the same”. We do not propose that this should mean that all distinctions should be made in all languages. For example, those Uralic languages without object conjugation should not be required to adopt the agreement tags of those that have it. But it should be possible to come up with principled and consistent guidelines for closed categories.

7 Universal dependencies

Universal dependencies is a large multi-language project [7] aiming at common tagset for part-of-speech, morphosyntactic features and dependency relations. We do not propose adopting the exact tagset of the universal dependency project. Most projects working on Uralic languages have been ongoing for many years and the tools that they create are used for more than just machine translation. What we find more important is to adopt, or make available tools based on a consistent theoretical background and consistent morphosyntactic description. This could form the basis of a kind of *universal* morphosyntactic interlingua for the Uralic languages. These tools do not have to replace the current tools, and may be automatically generated from them, but they must be consistent. A systematic mapping needs to be considered while developing. The national Uralic languages have specifications for universal dependencies [8, 9, 10]. But these specifications differ in unnecessary ways. For example, consider the annotation of ‘that house’ in the two treebanks for Finnish: Turku Dependency Treebank (TDT) and FinnTreeBank (FTB); and Hungarian:

	this		house
Finnish (TDT)	tämä$_{\text{PRON}}$		talo$_{\text{NOUN}}$
Finnish (FTB)	tämä$_{\text{DET}}$		talo$_{\text{NOUN}}$
Hungarian	az$_{\text{PRON}}$	a$_{\text{ART}}$	ház$_{\text{NOUN}}$

8 Concluding remarks

Rule-based machine translation provides a fascinating basis for exploring real linguistic differences between the Uralic languages. However, as we have shown, in current state-of-the-art tools, real linguistic differences are hidden behind a combination of incompatible tagsets and idiosyncratic traditional grammatical norms. We do not propose that the North Sámi adopt the Finnish norms or the Hungarians the Erzya norms, instead we propose developing a common morphological annotation scheme for the Uralic languages based on guidelines of the Universal dependencies project. It is not our aim for this to supercede national standards, but provide a common bridge between them to facilitate the cross-linguistic study and functional rule-based machine translation.

Acknowledgements

Heiki-Jaan Kaalep, Jack Rueter, László Tihany as well as the anonymous reviewers have all contributed to the language examples, the remaining mistakes are ours.

A Example of Universal dependencies for Uralic languages

Example is shown in table 4.

References

[1] Mikel L. Forcada, Mireia Ginestí Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Francis M. Tyers. Apertium: a free/open-source platform for rule-based machine translation platform. *Machine Translation*, 2010.

[2] Pim Otte and Francis M.Tyers. Rapid rule-based machine translation between dutch and afrikaans. In *Proceedings of the 15th conference of the European Association for Machine Translation, 30-31 May 2011, Leuven, Belgium*, pages 153–160, 2011.

[3] Jack Rueter. *Adnominal person in the morphological system of Erzya*. PhD thesis, 2010.

[4] Tommi A Pirinen. Omorfi–Free and open source morphological lexical database for Finnish. In *Nordic Conference of Computational Linguistics NODALIDA 2015*, page 313, 2015.

[5] V. Trón, A. Kornai, G. Gyepesi, L. Németh, P. Halácsy, and D. Varga. Hunmorph: open source word analysis. In *Proceedings of the Workshop on Software*, pages 77–85. Association for Computational Linguistics, 2005.

James	*ja*	*Mary*
PROPN	CONJ	PROPN
Number=Sing|Case=Nom		Number=Sing|Case=Nom
leaba	*gárdimis*	.
VERB	NOUN	PUNCT
Mood=Ind|Tense=Pres|Person=3|Number=Dual	Number=Sing|Case=Loc	
Джеймс	*марто*	
PROPN	CONJ	
Number=Sing|Case=Nom|Definite=Ind		
Марит	*садпиресэ-*	
PROPN	NOUN	
Number=Plur|Case=Nom|Definite=Ind	Case=Ine|Definite=Ind	
-ть	.	
VERB	PUNCT	
Mood=Ind|Tense=Pres|Pers[subj]=3|Number[subj]=Plur		
James	*ja*	*Mary*
PROPN	CONJ	PROPN
Number=Sing|Case=Nom		Number=Sing|Case=Nom
ovat	*puutarhassa*	.
VERB	NOUN	PUNCT
Mood=Ind|Tense=Pres|Person=3|Number=Plur	Number=Sing|Case=Ine	
James	*ja*	*Mary*
PROPN	CONJ	PROPN
Number=Sing|Case=Nom		Number=Sing|Case=Nom
on	*aias*	.
VERB	NOUN	PUNCT
Mood=Ind|Tense=Pres|Person=3|Number=Plur	Number=Sing|Case=Ine	
James	*és*	*Mary*
PROPN	CONJ	PROPN
Number=Sing|Case=Nom		Number=Sing|Case=Nom
kértben	.	
NOUN	PUNCT	
Number=Sing|Case=Ine		

Table 4: An example of applying universal part-of-speech tags and morphological features to the Uralic languages. Note how the massive differences in annotation are reduced to only the linguistically relevant compared to Table 1.

[6] Aarne Ranta. *Grammatical framework: Programming with multilingual grammars*. CSLI Publications, Center for the Study of Language and Information, 2011.

[7] Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. In *ACL (2)*, pages 92–97. Citeseer, 2013.

[8] Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. Universal dependencies for finnish. In *Nordic Conference of Computational Linguistics NODALIDA 2015*, page 163, 2015.

[9] Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, and Dage Särg. Estonian dependency treebank and its annotation scheme. In *Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 285–291, 2014.

[10] Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. Hungarian dependency treebank. In *LREC*, 2010.