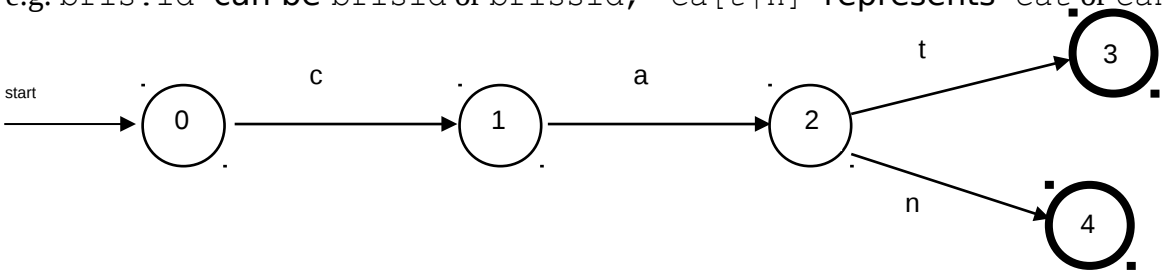
Th. Fransen, 11/10/17

A selection of issues with Finite-State Morphology for Old Irish verbs

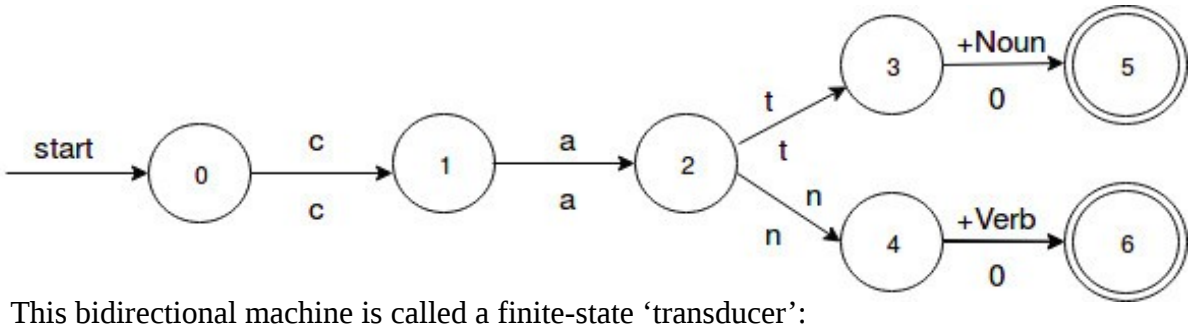
mainly for 'test cases' (see below), but with example inflections from other verbs to illustrate problems and challenges

Part A. What is finite-state morphology (in a nutshell)?

Finite-state machines are networks that can be generated (are mathematically equivalent to) regular expressions, abstract string matching patterns:
e.g. `bris?id` can be `brisid` or `brissid`; `ca[t|n]` represents `cat` or `can`:



State 3 and 4 are final, at that stage the 'machine' accepts the concatenated strings (=words) c-a-t and c-a-n.
An extension of this principle is a network that has symbols on *both* sides of the arcs, e.g.



This bidirectional machine is called a finite-state 'transducer':
The upper/lexical level reads: `cat+Noun`, mapped to `cat` on the surface level (`0` on the lower level is the empty string mapped to `+Noun`, and is 'ignored' in the transition, i.e. invisible). Similarly for the verb (string) `can`. For `cat` we could extend with a 7th state connected to state 5 with `+Plural` on the upper and `s` on the lower level giving `cats`, etc. etc.
This formalism has proven to be very efficient and elegant for rule-based morphology and phonology. It is inherently bidirectional: at all stages can we *analyse* (look up → what grammatical info belongs to `cat`) and *generate* (look down → what is the surface form of `cat+Noun+Plural`).
A rule-based finite-state approach generally consists of two ways of modelling word formation, reflected in two file formats/software tools that I use in conjunction with each other:
1. a **lexicon** with all the entries (e.g. verbs) that concatenates stems/roots with affixes/endings, leading to abstract strings on which are performed:
2. rewrite rules, to model non-concatenative morphophonemic processes to create surface/orthographical forms, using rules echoing phonological rewrite rules of the type
`n → m || _ p`, i.e. 'n becomes m before p', i.e. nasal assimilation, common in many languages.

By an operation called *composition*, we can combine ('collapse), in cascade, various transducers (bidirectional networks) representing a lexicon and (many) rules into one final transducer with the lexical string at the top, and the final surface string at the bottom, see page 8. (but, as said, we can start from the bottom and 'read' upwards as well). Now over to (mainly) Old Irish….

PART B: Ending variation in Old Irish

Overview of verb test cases employed in PhD thesis

Root	**Pres. 3sg deponent**	**Pres. 3sg active (indep.), eDIL headword**	**Present stem class**	**standard Modern Irish (impv. sg.)**	**Translation**
mol	molaithir	→ molaid	W1 (deponent)	mol	to praise
bris	x	bris(s)id	W2	bris	to break
(imbi-)tēg	x	imm·téit	S1	imigh	to go about
(inde-)fēt	x	in(d)·fét	S1	inis	to tell
(ess-)ber	x	as·beir	S1	abair	to say

In general about verbal endings listed in Stifter (in Ball/Mueller 2010: 90-1):
“Due to the quirks of Old Irish orthography there is much more spelling variation in the endings than can be indicated in these tables. In particular, the spellings of the vowels in the tables reflect those of the central examples of each class. There is, however, always the possibility that the endings appear after a consonant of the opposite quality. Furthermore, with endings of the shape -CV there is usually also a variant with a vowel before the consonant. The distribution of allomorphs depends on the syncope pattern.”

‘disagreement’ between consonant quality and endings: easy examples

deponent

pres. 3sg. deponent conj. *-ithir*
generates *molithir* ‘he praises’, root *mol-*, but also frequently attested (with vowel “congruence”) *mol**a**ithir*.

a-endings

The endings given by Stifter (2010: 91) only work for stems ending in a non-palatal consonant, e.g. subj. 1sg. **bris(s)-a → bris(s)-**e**a* ‘I may break’, 2sg. **bris(s)-ae → bris(s)-**e*** ‘you may break’.

This ending variation is predictable with *weak* stem classes (W1, W2), since these stems are either ‘non-palatal’ or ‘palatal’ throughout. I expect that variation seen in forms such as *molaithir* and *molither* can be handled by rewrite rules, perhaps in combination with implemented ‘knowledge’ about orthographical rules for vowels in unstressed syllables (see below).

Endings of the shape -CV (and vowel intrusion)

A vowel might be inserted before endings of the type -CV, e.g. s-future 2pl. *cich**e**-s-te* ‘you will step’ (ending *-ste* given by Stifter 2010: 92, with *s* future stem consonant). Other examples occur with the secondary endings.

Secondary endings (Stifter 2010: 94)

Inflection	**Secondary endings**
	(conj.)
1sg.	-(a)inn
2sg.	-th(e)a
3sg.	-ad, -ed
pass.	-th(a)e
1pl.	-m(a)is
2pl.	-th(a)e
3pl.	-t(a)is
pass.	-t(a)is

With the secondary future (conditional) of weak verbs taking the *f* stem consonant (*f*-future), we can expect forms such as pass. sg. *bris-f-**i**the*, 2pl., *bris-f-**i**the*,, with insertion of *i.*
A vowel is also inserted with the ending template -CVC and -C (not explicitly mentioned by Stifter):

-C:
e.g. pret. pass. 1sg. conj., zero ending: e.g. **·bris-th-0 → bris-**e**th* (or *bri-s-**e**d*)
-CVC
e.g. conditional 1pl. *bris-f-**i**mis*, 3pl., *bris-f-**i**tis.* Another example involving a secondary ending (but with a strong verb, taking an *s*-future) is *·eic-s-**i**tis* (from *in(d)·fét*, cited in eDIL, conditional 3pl., prototonic).
I'm not sure if the processes with a different ending template can be classified as the same rule, but the process in general seems to be one of vowel intrusion when a stem consonant and a consonant in the ending come together, here the future stem consonants *f* and *s* of the future formation, and *th* of the pret. passive.

<u>Compounds, strong verbs, irregularity and syncope: more challenging issues…</u>[1]
All compounds in our test set are S1.
Present tense S1

Inflection	**Pres III Ending (Stifter 2010: 92)**	
	abs.	**conj.**
1sg.	-u	-ø (+ u-inf.)
2sg.	-i	^{j}-ø (+ raising)
3sg.	-id	^{j}-ø
rel.	-as	-
1pl.	-mai	-am
1pl. rel.	-mae	-
2pl.	-the	-id
3pl.	-ait	-at
3pl. rel.	-tae	-

1This section contains a mix of different issues, already partly mentioned above, I'm just trying to come to terms with the variation encountered.

Present paradigm test cases, compounds (empty cell: not found in grammars/eDIL, yellow: not found, filled in based on *do·tét*, red: in disagreement with ending set given above. For *in(d)·fét* I could not find the present inflection table in grammars and eDIL and this is partly based on the entry *ad·fét* from eDIL (dil.ie/440) and Wiktionary, https://en.wiktionary.org/wiki/adf%C3%A9t))

	imm·téit		**in·fét**		**as·beir**	
	conj. (deut.)	**conj. (protot.)**	**conj. (deut.)**	**conj. (protot.)**	**conj. (deut.)**	**conj. (protot.)**
1sg.	imm·tíag	·imtheg	ad·fédiu, ad·fíadu, ad·féidim, deut. or protot.??)	inedim, indeidim, (eDIL, ipv 1sg indiad)	as·biur	·epur
2sg.	imm·téig	·imthig	ad·fédi	(·indid, eDIL, ipv 2sg)	as·bir	·epir
3sg.	imm·téit	·imthet	in(d)·fét, ad·féit, ad·fét, ad·fíad, asid·fét (with inf. pron.)		as·beir	·epir
rel.	imm**e**·t(h)ét		ib. (lenition of *f* not shown in OI orthography)		ib. (lenition of b not shown in OI orthography)	
1pl.	imm·tíagam	·imthigem	ad·fíadam, ad·fédem, asnda·fíadam (with inf. pron.)		as·beram	·eprem
1pl. rel.	imm**e**·t(h)íagam		ib. (lenition of *f* not shown in OI orthography)		ib. (lenition of *b* not shown in OI orthography)	
2pl.	imm·téit	·imthet	ad·fédid		as·berid	·eprid
3pl.	imm·tíagat	·imthiget	in·fíadat, ad·féthet, at·fétet (with inf. pron.)		as·berat	·epret
3pl. rel.	imm**e**·t(h)íagat		ib. (lenition of *f* not shown in OI orthography)		ib. (lenition of *b* not shown in OI orthography)	

We can see two stem patterns (allomorphs) in the prototonic stems of *as·beir*: *·epVr* and *·epr-*. This is the result of syncope (deletion of even syllables, except for the last one) after the addition of a third syllable (**e**ss-b*é*r-X → *epr-*).
We can see how endings differ when a preceding consonant is palatal. In the present indicative prototonic stem paradigm of *as·beir*, 'non-palatal' and 'palatal' endings differ according to vowel in the ending: the *e* of *ber* is a front vowel, which after syncope in the prototonic forms causes the *r* to become palatal. The *a* in the endings, for example, change into *e*, according to the orthographical rules of Old Irish: a vowel preceded by a palatal and followed by a non-palatal consonant in unstressed syllables of the type CVC, where V is always ə, is spelt *e*. The orthographical rules for vowels in *unstressed* (i.e. non-initial) syllables are the following:
C ə C → a (e.g. *as·ber**a**m*)
C^j ə C → e (e.g. *·epr**e**m*)
C ə C^j → (a)i (e.g. *mol(**a**)**i**thir, mol(**a**)**i**d)*
C^j ə C^j → i (e.g. *mol(a)ith**i**r*)

Automatising this system, i.e. generating the correct endings, requires knowledge of the quality of the final consonant of the stem (we don't have to be concerned about stressed or unstressed position and the orthographical rules concerning vowels: an ending is always non-initial, i.e. unstressed). The tricky part is that consonant quality may change due to syncope. Hard-coding a prototonic stem (as I've done in my current implementation) rather than deriving it on the basis of the root and 'calculate' the final consonant quality (involving rules such as _s+b_ → p, e.g. *ess-ber → epr*, lost vowel is *e*, *ergo* palatal *r*; and this is a fairly easy case!), entails that we have to mark explicitly whether the stem takes a 'non-palatal' or 'palatal' ending set (we don't know, after all, whether the *r* in *epr* is non-palatal or palatal since we haven't derived it, unless we code this. In my terms: which continuation lexicon? (I'll explain in person). I've made a start with dividing ending sets up in 'palatal' and 'non-palatal' in the current version of the transducer, but more rules are required to choose the right ending (or to be implemented differently, invariably catering for the 'opposite' stem-final consonant quality and the 'opposite' ending set?).

PART C: Stem variation

We already saw some variation with S1 verbs in this regard, e.g. *epir, epr- and* pres 3pl. *in·féthed*, *in·fétet* as opposed to *in·fíadat* etc.
This section goes more in details, focusing on the subjunctive.

VARIATION WITH THE S SUBJUNCTIVE
"The *s*-marker […] is subject to fluctuation in quality like the stem final of S1(a) presents." (McCone, EIV, p. 33)
The stem consonant in the present of S1(a) forms seems to be generally palatal with 2 and 3sg but plural forms vary*:* absolute *smai/smi, stae/ste, sait/sit*, cf. e.g. *ge-**s**-**mi*** , *ge-**s**-**tae**, ge-**ss**-**ait*** (1/2/3 pl. subj of *guidid* 'to pray') , *mái-**s**-**mi**, mái-**s**-**te,** má-**ss**-**ait*** (1/2/3 pl. subj of *maidid* 'to break')**,** *sé-**s**-**mai*** , *séi-**s**-**te**, sé-**ss**-**ait*** (1/2/3 pl. subj of *seinnid* 'to play') , *sle-**s**-**mi,** sle-**s**-**tae**, sle-**ss**-**ait*** (1/2/3 pl subj of *sligid* 'to fell'). ANY RULES HERE???
In the past subjunctive (pres. subj. stem + secondary endings) the stem final consonant quality and endings are somewhat more predictable, 1sg and 3sg. generally being 'palatal' and the rest of the paradigm 'non-palatal', e.g. 1sg. *imm·téi-**s**-**inn**, ·céi-**s**-**inn**, ·mái-**ss**-**inn*** (*imm·téit, cingid* 'to step' S1, *maidid* 'to break', S2.) and 3sg. *imm·téi-**s**-**ed**, ·céi-**s**-**ed**, ·mái-**ss**-**ed**,* but other person/number with *ía,* e.g. *imm·**tía**-**s**-**ta** (*2sg*), má-**s**-**mais*** (1pl), *cía-**s**-**tae*** (2pl), *imm·tía-**s**-**tais*** (3pl). However, see *saigid* attested past subj. 3sg *·sá-**ss**-ad* as well as *·sái-**ss**-**ed*** (Green, Old Irish V&V, p. 90), *claidid* (S1) *·clá-**ss**-**ainn*** (1sg). We are reminded of Stifter's observation above that endings may appear after a consonant of the opposite quality.

PART D: Potential computational implementation of stem and ending variation
(based on s-subjunctive examples)

cingid→s subjunctive stem *céis/cías*, implemented with triggers in the lexicon (`^X`), intermediary strings coded in that signify non-concatenative processes (mainly phonological rewrite rules) that we want to sort out after the initial 'blunt' concatenation of endings.
Challenge: cater for variation *éi/ía* depending on the ending by defining an underspecified lexical (underlying) stem vowel `^É` and by marking an ending that differs across verbs by using `^A` for a potential *a* (dependent on verb, e.g. *cíasm**a**i* as opposed to *gessm**i***). `^P` is a trigger to mark palatalisation of the final stem consonant (redundant here, as the rewriting of `É` to `éi` already caters for this); `^ZERO` is a null-ending. The trigger `^S` will be rewritten as *s* or *ss*, not relevant for the discussion here.

1sg `c^É^Su`
2sg `c^É^Si`
3sg `c^É^S^P`
rel. `c^É^S^ZERO`
1pl `c^É^Sm^Ai`
rel. `c^É^Smae`
2pl `c^É^St^Ae`
3pl `c^É^S^Ait`
rel. `c^É^Stae`

Optionally rewrite `^A` as `a`.

```
^A (->) a
```

Rewrite rule for *cías*-:
The right context symbols aim at cases with a non-palatal stem consonant *s*, among others `a` being mapped ('rewritten') from `^A`):
```
[ %^É -> {ía} || _ Cons  (Cons) (Cons) [%^R | u | %^ZERO  | a ] ];
```

Rewrite rule for *céis*-:
invoking (among others) `^A` in the right context which has **not** been mapped to `a` and is followed by `e` (2pl.)
```
[ %^É -> {éi} || _ Cons (Cons) (Cons) [%^LV | i |%^P | %^A e] ];
```

Delete all `^A`'s (and other triggers)
```
 ^A -> 0
```

e.g. 1pl `c^É^Sm^Ai` → *cías(s)mai*
e.g. 2pl. `c^É^St^Ae` → *céis(s)te*

The residue reflects cases where either `^A` or `a` (from `^A`) have not triggered `É` to be rewritten, e.g. `c^É^Smi`, which should neither become *cíasmi* (violates orthographical rules) nor *céismi* (not expected or attested (?)). In other words, we avoid the production of illegal and unattested forms.

A few examples of the output for subjunctives from my transducer with this idea partly implemented below, to show you how the upper level (lexical) tags and the final surface string map to each other. Various rewrite rules, incl. the ones above, produce intermediary strings that are not visible in the output. I will show / explain those rules in person, as well as elucidating the tags and how I can quickly do this mapping between upper and lower level...

+CONJPART+NEG+INTERR+AUGM+OBJ+3P+SG+MASC+cing+ROOT+S1+SIMPLE+SUBJ+CONJ+1P+PL
innádracíasam
+ind+PREVERB+fēt+ROOT+S1+DEUTEROT+PRES+SUBJ+CONJ+3P+PL
infesat
+imbi+PREVERB+tēg+ROOT+S1+DEUTEROT+PRES+SUBJ+CONJ+2P+PL
immtéissid

The above system can be used for stem and ending variation in general, e.g. subj. 1pl. *ge-ss-mi (guidid)* or *sle-ss-mi (sligid)*, with *-mi* instead of *-mai* depending on the vowel in the (in this case subjunctive) stem. Signifying a 'non-palatal' ending with `^A` facilitates the creation of correct endings and/or variation and can be extended to other tenses/moods, e.g. *mol(a)id*.

A more general problem, of course, is that it is to a large extent unpredictable which present stem class takes which non-present stem class. In our example above S1 and S2 verbs are shown forming an s-subjunctive. However, S verbs may take other formations; S2 *daimid* 'to permit', for example, takes an ā-subjunctive formation. These non-present stems need to be individually listed (hard-coded) in the lexicon (e.g. present *imm·téit*, but subj. *imm·téis/imm·tías* (possibly `immt^És` analogous to the preceding) and this is the way I try to overcome the general stem variation problem in the current version of the transducer.

PART E: More broadly, my research thesis, not terribly important for your purposes ;-)

Mapping Old : Modern Irish

Old Irish	**Middle Irish**	**Classical Modern Irish**	**Modern Irish**
molaithir/molaid, ·molathar/·mola	molaid, ·mol(ann)	molaidh, mol(ann)	molann
brisid, ·bris	brisid, ·bris(enn)	brisidh, ·bris(eann)	briseann
imm·téit, ·imthet	imthigid, ·imthig(enn)	imthighidh, imthigh(eann)	im(th)igheann / im(th)idheann > imíonn
ind·fét, ·indet	indisid, ·indis(enn)	innisidh, ·innsidh(eann)	in(n)(i)seann / in(n)sidheann / in(n)isigheann >insíonn
as·beir, ·epir	at·beir, ·epir(?)	a·deir (?)	deir(eann)
		abair (?)	abair

Mappings from Old Irish compound stems to Middle Irish compound stems are often unpredictable yet crucial in linking Old Irish forms to later forms. Subsequent changes (i.e. beyond Middle Irish) are mainly (although not exclusively) in the realm of orthography. Therefore, somehow in the computational architecture, mappings need to take place between Old and Middle Irish. If we want

to implement these mappings in the finite-state transducer for Old Irish verbs, we need a “Middle Irish stem mapping” component, taking the form of e.g.

imbi+PREVERB+tēg+ROOT+S1+DEUTEROT+PRES+IND+3P+SG
immtéit

.o. (composition)

immtéit
imthigid

=

imbi+PREVERB+tēg+ROOT+S1+DEUTEROT+PRES+IND+3P+SG
imthigid

This model is too simplistic and cannot generate/analyse many of the desired Middle Irish forms on the basis of Old Irish compound verb forms. One of the complications is stem variation outside the 3 sg. present. Even in the Old Irish present tense, we find stem alternation, namely, with strong verbs. The 3pl. present indicative of imm·téit is imm·tíagat, with the (conjunct) ending -at. As such, we can identify another stem here, imm-tíag. The following table shows a selection of different inflections for *imm-téit*

Old Irish		Middle Irish (univerbation, simple verb)
pres 3sg. deut.	imm·téit	imthigid
pres 3sg. protot.	·imthet	·imthig(enn)
pres 3pl. deut.	imm·tíagat	immthigit
pres 3pl. protot.	·imthiget	·imthiget
pret 3sg. deut.	imm·luid	immthigis
pret 3sg. protot.	·imlaid	·imthig

And of course, as can be seen in the table above, with compounds we invariably deal with a second set of stems, i.e. the prototonic stems appearing in dependent position (e.g. *imthet*), which may also take a different shape across the paradigm for a single tense/mood. This effectively means that we need one-to-more mappings (from the viewpoint of Middle Irish), e.g.

imm·téit:imthig
imthet:imthig
imm·tíag:imthig
imm·luid:imthig
imlaid:imthig
etc.

New Middle Irish stems (e.g. *imthig*) are inflected according to the weak verb inflection of Old Irish. In other words, *imthig* now effectively corresponds to W1 *bris*, taking the predictable ‘patatal’ weak stem ending set, which undergoes slight changes in terms of the present, e.g. 3sg present conjunct ending -0 indicative of the Old Irish strong verbs, *ní imthig* (**imthig-i*).