Treex Cheat Sheet

http://ufal.mff.cuni.cz/treex/

treex options

Basic options

-h, -v	help, version
-q	only fatal errors reported to stderr
-e	error_level: ALL, DEBUG, INFO, WARN, FATAL
-L	-Lcs ∼ Util::SetGlobal language=cs, default=all
-S	-Smst ∼ Util::SetGlobal selector=mst, default=''
-d	just dump scenario to stdout and exit
-t	with -Lxy ∼ Read::Sentences W2A::XY::Tokenize
-s	append Write::Treex ∼ modify treex files in-situ
--	-- *.treex.gz ∼ Read::Treex -- '!dir1,dir2/*/*.txt' ∼ Read::Text

Parallel exectution options

-p, -j	parallel execution, # jobs, e.g. -pj5
--local	run locally (multi-core), otherwise SGE
--mem	-m=5g ∼ qsub -hard -l mem_free=5g -l act_mem_free=5g -l h_vmem=5g
--priority	integer in the range -1023 to 0, default=-100
--survive	don't kill all jobs when one crashes
--qsub	additional parameters for qsub, e.g. --qsub="-q *@p*,*@s*" ∼ use machines p* & s* --qsub="-m abe -M x@y.cz" (abort, begin, end)
--name	prefix of submitted jobs (qsub -N)
--workdir	temp directory, default=001-cluster run, 002...
--cleanup	delete the temporary workdir, default=false
-E	forward_error_level (to the main stderr)
--cache	--cache 1,9 ∼ use memcached (1GB+9GB)

Config

See http://ufal.mff.cuni.cz/treex/install.html

ÚFAL specific (legacy)

```
cd; svn co https://svn.ms.mff.cuni.cz/svn/tectomt_devel/trunk/treex treex
ln -s /net/projects/tectomt_shared share
perlbrew switch perl-5.18.2

# in $HOME/.bashrc
source /net/work/projects/perlbrew/init
eval "$(bash-complete setup)"
export TMT_ROOT=$HOME
export PATH="${TMT_ROOT%/}/treex/bin:$PATH" # for treex and ttred
export PERL5LIB="${TMT_ROOT%/}/treex/lib:$PERL5LIB"

# in $HOME/.treex/config.yaml
---
resource_path:
  - /net/projects/tectomt_shared
share_dir: /net/projects/tectomt_shared
share_url: http://ufallab.ms.mff.cuni.cz/tectomt/share
tmp_dir: /COMP.TMP
tred_dir: /home/cinkova/tred/tred
pml_schema_dir: /home/popel/treex/lib/Treex/Core/share/tred_extension/treex/resources
tred_extension_dir: /home/popel/treex/lib/Treex/Core/share/tred_extension
```

Readers

Common parameters:
from – list of filenames (separated by spaces or commas)
  from=- read stdin (default for text readers)
  from=a.treex,dir/b.treex
  from='!{a,b}/0?/*.treex.gz' use ! for wildcards
  from='@filelist1,@filelist2' use @ for file lists
language – default='und'
selector – default='' (empty)
file_stem – name of the loaded documents (cf. aligned readers)
skip_finished – resume previous unfinished treex run
  skip_finished={indir/(.+).conll$}{outdir/$1.treex.gz}
For "text" readers:
encoding – default=utf8
lines_per_doc – split input file into more documents (default=0)
merge_files – default=0

Treex::Block::Read::*
- Treex – reads *.treex, *.treex.gz, *.streex
- Text – unsegmented plain text to $zone->text
- Sentences skip_empty=0 – one plain-text sentence per line
- CoNLLX lines_per_doc=9 – 9 sentences (!) per document
  columns: ord form lemma cpos pos feats head deprel
- CoNLL2009 – columns: ord form lemma plemma pos ppos feats pfeats head phead deprel pdeprel semantic_feats
- AttributeSentences attributes=form,lemma,tag,afun,parent
  one sentence per line (1-based parent index), default: layer=a
  default regex separators: separator=' ' attr_sep='\|'
- PDT schema_dir=resources/ t_layer=1 reads *.t.gz (or *.a.gz)
- PennMrg reads PennTreebank *.mrg constituency trees

multilingual (aligned) readers:
- SentencesTSV langs=en,cs tab-separated sentence tuples per line
- AlignedSentences en=en1.txt,en2.txt cs_ref=cs1.txt,cs2.txt
  parameters in form language(_selector)? instead of from
- AlignedCoNLL en='!e*.conll.gz' de='!d*.conll.gz' (CoNLL2009)

Writers

Common parameters:
to – space or comma separated output filenames
  to=- (stdout) is default for "text" writers,
  use to=. to use the document name (default for other writers)
substitute={dir(\d+)/file(\d+).treex}{f\1-\2.streex}i
compress=1 – *.gz
encoding=utf8 – default for "text" writers

Treex::Block::Write::*
- Treex (for fast *.streex use storable=1)
- Text – prints $doczone->text
- Sentences – prints $zone->sentence
- CoNLLX pos_attribute=conll/pos ... see POD
- AttributeSentences – see POD & Write::LayerAttributes::*
  layer=a attributes=form,lemma,tag tab-separated f|l|t
  attributes='tag parent->lemma' separator='\n' attr_sep='\t'
- PDT – saves *.w,*.m,*.a,*.t files
- PennMrg – PennTreebank *.mrg constituency trees
- TreesTXT sents=1 afuns=1 – legible dependency trees (ASCII-art)

Common Block Parameters

language – comma separated list of language codes to process
  default=all ∼ process zones with any language
selector – comma separated list of selectors (default='')
  e.g. language=en,cs selector='orig,,new' (6 zones)
select_bundles=1-4,6,8-12 default=0 ∼ process all bundles
report_progress=1 ∼ print which bundle is processed (via log_info)
if_missing_tree – what to do if process_[atnp]tree finds no tree in zone
  options: fatal (default), warn, ignore, create
if_missing_zone – dtto for process_zone
if_missing_bundles – dtto for no bundles in a document (~~create~~)

Useful Blocks

- Util::SetGlobal language=en my_tool_model=dir/my.dat
- Util::Eval – apply ad hoc Perl code on doc/bundle/.../node
  document='print $document->full_filename."\n"'
  doc='say $.full_filename'
  zone='$zone->remove_tree("t") if $zone->has_tree("t")'
  ttree='say $.language, "\t", scalar $.get_children()'
  anode='say $.form if $.parent->precedes($anode)'
- Util::Find – print get_address of matching [atnp](node|tree)
  tnode='$tnode->gram_gender eq "fem"'
  max_nodes_per_tree=1 default=0 ∼ all nodes
  on_error – fatal (default), warn, ignore
  treex Util::Find anode='$.is_member' -- *.treex | ttred -l-
- Util::PMLTQ query='t-node [t_lemma = "být"];'
  query_file – read the queries from the file
  one_per_match=1 ∼ print only one node address per match
  action Perl code to be executed on matching @nodes (or $node)
- Filter::SentenceNumber nums=1,3 invert=1 – delete bundles 1, 3
- A2A::BackupTree to_selector=my_backup – copy a-trees
  to_language – default (empty) means the same language
  to_selector – default (empty) means the empty selector
  flatted=1 ∼ target (backup) trees flat and is_member=0
  align=1 ∼ add alignment links of type copy
  keep_alignment_links=1 ∼ incoming and outgoing links preserved

Schematic Treex::Core::Block

```
sub process_document {
    my ($self, $document) = @_;
    foreach my $bundle ( $document->get_bundles() ) {
        $self->process_bundle($bundle);
    }
}

sub process_bundle {
    my ($self, $bundle) = @_;
    my $zone = $bundle->get_zone( $self->language, $self->selector );
    return $self->process_zone($zone);
}

sub process_zone {
    my ($self, $zone) = @_;
    if ($zone->has_atree()){
        $self->process_atree($zone->get_atree());
    }
    # similarily for other layers (t, n, p),
    # but currently only one process_[atnp](tree|node) method can be overriden
}

sub process_atree {
    my ($self, $atree) = @_;
    foreach my $node ( $tree->get_descendants() ) {
            $self->process_anode($node);
        }
    }
}

sub process_anode {
    my ( $self, $anode ) = @_;
}
```