GREC Corpus Evaluation Plan

Anthony Gentile
Lisa Gress

4/29/2013

Initial Corpus Coverage

◆ 240 MEDLINE abstracts totaling 1500 sentences.
◆ Using ACE (thanks Woodley) to find valid parses from the ERG.
◆ Results using Condor with 8GB:
  ◆ Valid parses: 1196 (79.7333%)
  ◆ No parse/Insufficient RAM: 304 (20.2666%)

What we are working with

The ompB operon encodes OmpR and EnvZ, two proteins that are necessary for the expression and osmoregulation of the OmpF and OmpC porins in Escherichia coli.

Example MRS

INDEX: e2 [e SF: prop]
RELS: < [_the_q_rel<0:3> LBL: h4 ARG0: x3 [x PERS: 3 NUM: sg] RSTR: h5 BODY: h6]
 [compound_rel<4:15> LBL: h7 ARG0: e8 [e SF: prop TENSE: untensed MOOD: indicative PROG: - PERF: -] ARG1: x3 ARG2: x9 [x IND: +]]
 [proper_q_rel<4:8> LBL: h10 ARG0: x9 RSTR: h11 BODY: h12]
 [named_rel<4:8> LBL: h13 CARG: "ompB" ARG0: x9]
 ["_operon/NN_u_unknown_rel"<9:15> LBL: h7 ARG0: x3]
 ["_encode_v_1_rel"<16:23> LBL: h15 ARG0: e16 [e SF: prop TENSE: pres MOOD: indicative PROG: - PERF: -] ARG1: x3 ARG2: x17 [x PERS: 3 NUM: pl]]
 [udef_q_rel<24:38> LBL: h18 ARG0: x17 RSTR: h19 BODY: h20]
 [proper_q_rel<24:28> LBL: h21 ARG0: x22 [x PERS: 3 NUM: sg IND: +] RSTR: h23 BODY: h24]
 [named_rel<24:28> LBL: h25 CARG: "OmpR" ARG0: x22]
 [_and_c_rel<29:32> LBL: h27 ARG0: x17 L-INDEX: x22 R-INDEX: x28 [x PERS: 3 NUM: sg IND: +]]
 [proper_q_rel<33:38> LBL: h29 ARG0: x28 RSTR: h30 BODY: h31]
 [named_rel<33:38> LBL: h32 CARG: "EnvZ" ARG0: x28]
 [subord_rel<39:157> LBL: h1 ARG0: e34 [e SF: prop TENSE: untensed MOOD: indicative PROG: - PERF: -] ARG1: h35 ARG2: h36]
 [udef_q_rel<39:42> LBL: h37 ARG0: x38 [x PERS: 3 NUM: pl] RSTR: h39 BODY: h40]
 [card_rel<39:42> LBL: h41 CARG: "2" ARG0: e43 [e SF: prop TENSE: untensed MOOD: indicative] ARG1: x38]
 ["_protein_n_1_rel"<43:51> LBL: h41 ARG0: x38]
 ["_necessary_a_for_rel"<61:70> LBL: h41 ARG0: e44 [e SF: prop TENSE: pres MOOD: indicative PROG: - PERF: -] ARG1: x38 ARG2: x45 [x PERS: 3]]
 [_the_q_rel<75:78> LBL: h46 ARG0: x45 RSTR: h47 BODY: h48]
 [udef_q_rel<79:89> LBL: h49 ARG0: x50 [x PERS: 3 NUM: sg] RSTR: h51 BODY: h52]
 ["_expression_n_1_rel"<79:89> LBL: h53 ARG0: x50]
 [udef_q_rel<90:108> LBL: h54 ARG0: x55 [x PERS: 3 NUM: sg] RSTR: h56 BODY: h57]
 [_and_c_rel<90:93> LBL: h58 ARG0: x45 L-INDEX: x50 R-INDEX: x55]
 ["_osmoregulation/NN_u_unknown_rel"<94:108> LBL: h59 ARG0: x55]
 [_of_p_rel<109:111> LBL: h58 ARG0: e60 [e SF: prop] ARG1: x45 ARG2: x61 [x PERS: 3 NUM: pl]]
 [_the_q_rel<112:115> LBL: h62 ARG0: x61 RSTR: h63 BODY: h64]
 [compound_rel<116:136> LBL: h65 ARG0: e66 [e SF: prop TENSE: untensed MOOD: indicative PROG: - PERF: -] ARG1: x61 ARG2: x67 [x PERS: 3 NUM: pl]]
 [udef_q_rel<116:129> LBL: h68 ARG0: x67 RSTR: h69 BODY: h70]
 [proper_q_rel<116:120> LBL: h71 ARG0: x72 [x IND: +] RSTR: h73 BODY: h74]
 [named_rel<116:120> LBL: h75 CARG: "OmpF" ARG0: x72]
 [_and_c_rel<121:124> LBL: h77 ARG0: x67 L-INDEX: x72 R-INDEX: x78 [x IND: +]]
 [proper_q_rel<125:129> LBL: h79 ARG0: x78 RSTR: h80 BODY: h81]
 [named_rel<125:129> LBL: h82 CARG: "OmpC" ARG0: x78]
 ["_porins/NNS_u_unknown_rel"<130:136> LBL: h65 ARG0: x61]
 [_in_p_rel<137:139> LBL: h84 ARG0: e85 [e SF: prop TENSE: untensed MOOD: indicative] ARG1: x38 ARG2: x86 [x PERS: 3 NUM: sg]]
 [udef_q_rel<140:157> LBL: h87 ARG0: x86 RSTR: h88 BODY: h89]
 [compound_rel<140:157> LBL: h90 ARG0: e91 [e SF: prop TENSE: untensed MOOD: indicative PROG: - PERF: -] ARG1: x86 ARG2: x92 [x IND: +]]
 [proper_q_rel<140:151> LBL: h93 ARG0: x92 RSTR: h94 BODY: h95]
 [named_rel<140:151> LBL: h96 CARG: "Escherichia" ARG0: x92]
 ["_coli/NN_u_unknown_rel"<152:157> LBL: h90 ARG0: x86] >

GREC Annotation

```xml
<Abstract>
  <AbstractText>
    <sentence id="S3"><term sem="Operon" id="T6" lex="The_ompB_operon">The ompB operon</term> encodes <term sem="Protein"
id="T7" lex="OmpR_and_EnvZ">OmpR and EnvZ</term>, two proteins that are necessary for the expression and osmoregulation of <term
sem="Protein" id="T8" lex="the_OmpF_and_OmpC_porins">the OmpF and OmpC porins</term> <term sem="SPAN" id="T9"
lex="in_Escherichia_coli">in <term sem="Wild_Type_Bacteria" id="T10" lex="Escherichia_coli">Escherichia coli</term></term>.
    </sentence>
    <event id="E2">
      <type class="GRE" />
      <Agent idref="T6" />
      <Theme idref="T7" />
      <clue>The ompB operon <clueType>encodes</clueType> OmpR and EnvZ, two proteins that are necessary for the expression and
osmoregulation of the OmpF and OmpC porins in Escherichia coli.</clue>
    </event>
    <event id="E3">
      <type class="Gene_Expression" />
      <Theme idref="T8" />
      <Location idref="T9" />
      <clue>The ompB operon encodes OmpR and EnvZ, two proteins that are necessary for the <clueType>expression</clueType> and
osmoregulation of the OmpF and OmpC porins in Escherichia coli.</clue>
    </event>
    <event id="E4">
      <type class="Regulation" />
      <Theme idref="T8" />
      <Location idref="T9" />
      <clue>The ompB operon encodes OmpR and EnvZ, two proteins that are necessary for the expression and
<clueType>osmoregulation</clueType> of the OmpF and OmpC porins in Escherichia coli.</clue>
    </event>
    …
    …
  </AbstractText>
</Abstract>
```

Mapping Process

◆ For each sentence, find semantic role identifiers for each GREC corpus sentences:
  - ◆ <Location idref="T9">
  - ◆ <term sem="SPAN" id="T9" lex="in_Escherichia_coli">in <term sem="Wild_Type_Bacteria" id="T10" lex="Escherichia_coli">Escherichia coli</term></term>

◆ Map this semantic role to the corresponding MRS RELS ARGN for the given term.

```
[ _in_p_rel<137:139>
            LBL: h84
            ARG0: e85
            [ e
                        SF: prop
                        TENSE: untensed
                        MOOD: indicative
            ]
            ARG1: x38
            ARG2: x86
            [ x
                        PERS: 3
                        NUM: sg
            ]
]
[ udef_q_rel<140:157>
            LBL: h87
            ARG0: x86
            RSTR: h88
            BODY: h89
]
[ compound_rel<140:157>
            LBL: h90
            ARG0: e91
            [ e
                        SF: prop
                        TENSE: untensed
                        MOOD: indicative
                        PROG: - PERF: -
            ]
            ARG1: x86
            ARG2: x92
            [ x
                        IND: +
            ]
]
[ proper_q_rel<140:151>
            LBL: h93
            ARG0: x92
            RSTR: h94
            BODY: h95
]
[ named_rel<140:151>
            LBL: h96
            CARG: "Escherichia"
            ARG0: x92
]
[ "_coli/NN_u_unknown_rel"<152:157>
            LBL: h90
            ARG0: x86
]
```