**Prodigy project: Work packages**

<table>
<thead>
<tr>
<th>Month</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-3</td>
<td>-6</td>
<td>-12</td>
</tr>
<tr>
<td>Data creation</td>
<td>WP 3</td>
<td>WP 4</td>
<td></td>
</tr>
<tr>
<td>Building generators</td>
<td>WP 2</td>
<td>WP 5</td>
<td>WP 8</td>
</tr>
<tr>
<td>Representation</td>
<td>WP 7</td>
<td></td>
<td>WP 10</td>
</tr>
<tr>
<td>Probabilistic control</td>
<td>WP 6</td>
<td>WP 9</td>
<td>WP 12</td>
</tr>
<tr>
<td>Evaluation</td>
<td>WP 11</td>
<td>WP 11</td>
<td>WP 13</td>
</tr>
<tr>
<td>Software</td>
<td>WP 13</td>
<td>WP 13</td>
<td>WP 13</td>
</tr>
</tbody>
</table>

**WP 1** Adapt ILEX and BabyTalk corpora of paired inputs/outputs for data-to-text NLG (see Case for Support, Section 2.2).

*Deliverables:* ILEX and BabyTalk corpora, adapted for Prodigy purposes.

**WP 2** Build baseline generators for ILEX and BabyTalk corpora. Train using pCRU techniques, generate output texts for test set inputs.

*Deliverables:* domain-specific pCRU baseline generators for ILEX, BabyTalk; sets of test set outputs.

**WP 3** (PI) Create Prodigy Recipes corpus: harvest large collection of recipes from web; sort into recipe types and tidy up. Develop method for automatically extracting an input representation from recipe texts and list of ingredients (see Section 2.2); incorporate resulting input content representations into corpus.

*Deliverable:* Prodigy Recipes corpus of paired inputs and output sets.

**WP 4** (PI) Create Prodigy Census corpus: extract statistics from census, design interactive website for paraphrasing task, recruit, train and supervise subjects in paraphrasing task. Combine census statistics and paraphrases into corpus (see Section 2.2).

*Deliverable:* Prodigy Census corpus of paired inputs and output sets.

**WP 5** Build baseline generators for Prodigy Recipes and Prodigy Census corpora. Train using pCRU techniques, generate output texts for test set inputs.

*Deliverables:* domain-specific pCRU baseline generators for Prodigy Recipes and Prodigy Census corpora; sets of test set outputs.

**WP 6** Evaluation I: using automatic and human-based methods (see Section 2.5), evaluate test set outputs for all five pCRU baseline generators (including the existing SUMTIME generator), and outputs from corresponding existing hand-crafted generators. This will take longer in the first round than the second and third, because of evaluation infrastructure creation (websites, questionnaires, experimental design, etc.).

**WP 7** Create content representation formalism for inputs to data-to-text NLG that is sharable between all Prodigy domains, conducting overview of relevant existing methods and consulting panel of advisors (see Section 2.3).

*Deliverable:* Prodigy input representation formalism.

**WP 8** Reimplement pCRU generators for SUMTIME, ILEX, BabyTalk, Prodigy Recipes and Prodigy Census, this time using new input representation formalism, and encoding generation spaces with maximal overlap; generate test set outputs.

*Deliverables:* five generators using same domain-independent input representation formalism for the five domains; sets of test set outputs.

**WP 9** Evaluation II: as Evaluation 1, for new test set outputs.

**WP 10** Investigate alternative techniques for training and using probabilistic model; investigate alternative probabilistic models, and correspondingly different ways of encoding generation spaces; continuously test on the five Prodigy corpora, assessing progress with automatic evaluation.

*Deliverables:* final set of best performing techniques for probabilistic modelling for NLG, alpha version of Prodigy software package for building probabilistic generators.

**WP 11** Reimplement SUMTIME, ILEX, BabyTalk, Prodigy Recipes and Prodigy Census generators using final technology.

*Deliverables:* final generators for the five Prodigy domains.


**WP 13** (PI) Final testing of Prodigy software and corpora, packaging and release.

*Deliverables:* final versions of Prodigy Recipes and Prodigy Census corpora, and of Prodigy software for building probabilistic NLG systems, for release to researchers, under appropriate non-commercial use license.