Language Technology: Research and Development

R&D Projects – From Proposal to Implementation

Sara Stymne

Uppsala University
Department of Linguistics and Philology
sara.stymne@lingfil.uu.se

Partially based on slides by Joakim Nivre and Samuel Douglas

R&D Projects

Research and development is often organized into projects

- Time-limited
- One-time effort
- Specific goals
- Separate budget
- Separate organization

Projects vary in scope and size

- Term paper (1 person, 240 hours)
- EU FP project (15–20 sites, 6–10 MEUR)

Life Cycle of a (Funded) Project

Pre-grant activities:

- Explore research opportunities
- Write and submit research proposal
- Sign research contract

Post-grant activities:

- Start up: mobilize project resources
- Manage research activities
- Close down: report project outcome

Research Funding in Sweden – Government

Direct grants to universities (16 BSEK)

- Basic funding for research and graduate education

Research councils and agencies (9.5 BSEK)

- Swedish Research Council (VR)
- Environment, Agricultural Sciences and Spatial Planning (Formas)
- Health, Working Life and Welfare (FORTE)
- Innovation Systems (VINNOVA)

Note that the exact numbers are a few years old

Research Funding in Sweden – Other

Public research foundations (2.5 BSEK)

- Bank of Sweden Tercentenary Foundation (RJ)
- Swedish Foundation for Strategic Research (SSF)
- Knowledge Foundation (KK)

Other Swedish non-profit research foundations (3.5 BSEK)

- Knut and Alice Wallenberg Foundation
- Swedish Cancer Society

Industrial and business research (100 BSEK, 1 BSEK to universities)

Research Funding in the European Union

Horizon Europe (2021–2027) (95.5 BEUR)

1. Excellent science (25 BEUR)
   - European Research Council (16 BEUR)
   - Marie Curie (6.6 BEUR)
   - Research Infrastructures (2.4 BEUR)
2. Global challenges and industrial competitiveness (53.5 BEUR)
   - Health, food, culture, digital, climate, civil security, ...
   - Joint research centre
3. Innovative Europe (13.6 BEUR)
   - European innovation council and ecosystems, European institute of innovation and technology
4. Widening and Euratom (5.4 BEUR)

Planning a Project

Research question

- What are you going to find out?

Previous work

- What do we know already?

Approach

- How are you going to find out?

Time plan

- When are you going to do what?

Research Questions

A research question is a clear, focused, relevant, and interesting question around which we center our research.

- Clear: Stated concisely using precise terminology
- Focused: Not too broad (nor too limited)
- Relevant: Has a bearing on the research topic
- Interesting: Provides substantial new information if answered

Above all, questions should be researchable.

Research Questions – Good or Bad?

1. Does global warming affect parsing accuracy?
2. Do multiword expressions affect parsing accuracy?
3. How do multiword expressions affect parsing accuracy?
4. How do light verbs affect parsing accuracy?
5. Are light verbs harder to parse than other verbs?
6. What can we do to improve parsing accuracy for light verbs?
7. Can valency info improve parsing accuracy for light verbs?
8. What is the F_1 of the Berkeley parser on light verbs in PTB?
9. How do you install the Berkeley parser on your laptop?

Previous Work

- Why?
  - Scientific research should result in new knowledge
  - We make progress by building on previous results

    "If I have seen further it is by standing on the shoulders of giants." (Newton)
- How?
  - Find literature using a focused search (internet, library)
  - Manage the literature in a database (references, annotations)
  - Use the literature in your own work (context, motivation)
- Tips and tricks:
  - Start with handbook or survey articles if available
  - Use the snowball method (references of references)
  - Use citation statistics (with caution)

Useful Resources and Tools

- The ACL Anthology (https://aclanthology.info)
  - Repository of (currently) over 46,000 scientific papers
  - Searchable using general or specialized search engines
  - Full text articles (PDF) and bibliographic references (BibTeX)
- University library (http://ub.uu.se)
  - Databases (Web of Science, ScienceDirect, Google Scholar)
  - Journals and books (printed and electronic)
- Reference management software
  - BibTeX (used with LaTeX) – de facto standard in LT research
  - EndNote (widely used with MS Word)

Approach

- Theory:
  - Theoretical framework (concepts, definitions)
  - Refinement of research questions
- Method:
  - How can we answer the research question?
  - What theoretical results do we need (and how to prove them)?
  - What empirical data do we need (and how to get them)?
  - How do we analyze the results?
- Approach has to fit research questions

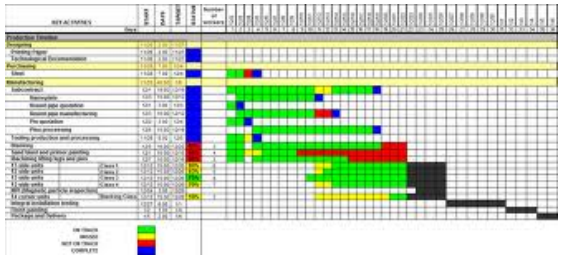
Approach – An Example

- Research question:
  - Are light verbs harder to parse than other verbs?
- Theory:
  - Parsing framework
  - Definitions (light verbs, other verbs)
- Method:
  - Data selection (sampling, annotation)
  - Evaluation metrics for verb-specific accuracy
  - Experimental setup (systems, data splits, tuning)
  - Hypothesis testing (statistical tests)
  - Error analysis (quantitative, qualitative)

Designing Experiments

- Identify variables:
  - Independent variable – manipulated by researcher
  - Dependent variable – measured by researcher
  - Control variable – kept constant by researcher
- Select data:
  - Avoid bias in data selection
  - Distinguish training, development and test data
- Design measurements and analysis:
  - Use appropriate metrics
  - Use a reasonable baseline
  - Repeat measurements if needed
  - Use appropriate statistical tests
  - Check for alternative explanations

Time Plan



- Devise a project plan:
  1. Identify tasks and subtasks
  2. Identify dependencies between tasks
  3. Order tasks and make time estimates
  4. Set up milestones and contingency plans
- Words of wisdom:
  1. Keep it simple!
  2. Keep deadlines deadly!
  3. Multiply all time estimates by three!

Writing a Project Proposal

- Scientific part:
  1. Introduction (research questions, motivation)
  2. Background (previous work, current issues)
  3. Project description (theory, method, time plan)
  4. Expected results (significance)
- Administrative part:
  1. Organization and management
  2. Deliverables and milestones
  3. Budget
  4. Participants' qualifications (CV, publications)

VR Guidelines (Research Plan)

- Purpose and aims
  - Present the overall purpose and specific aims of the project.
- Survey of the field
  - Summarize previous research with key references.
- Project description
  - Give a summary of the project describing its theory, methods, time plan, and implementation.
- Significance and scientific novelty
  - Describe short-term and long-term significance of the project.
- Preliminary and previous results
  - Describe pilot studies that support the feasibility of the project.

VR Guidelines (Research Plan)

- Purpose and aims
  - Present the overall purpose and specific aims of the project.
- Survey of the field
  - Summarize previous research with key references.
- Project description
  - Give a summary of the project describing its theory, methods, time plan, and implementation.
- Significance and scientific novelty
  - Describe short-term and long-term significance of the project.
- Preliminary and previous results
  - Describe pilot studies that support the feasibility of the project.

Hints for Writing a Project Proposal

- Content:
  - State research questions clearly and concisely from the start
  - Use background to motivate research questions
  - Be as specific as possible about theory and method
  - Avoid unnecessary details – convey the big picture
  - Make sure to follow the guidelines closely
- Form:
  - Use exact terminology (but avoid obscure technical jargon)
  - Use correct grammar and spelling (but keep it simple)
  - Use concrete examples to exemplify abstract concepts
  - Use graphical illustrations when appropriate
  - Respect page limits with reasonable margins and font sizes

Popular Scientific Writing

- Write so that the general public can understand (target your grandma!)
- Assume no knowledge of NLP
- Avoid terms, or explain them if they are needed
- Use simple language
- Focus on the most important points
- Avoid technical details (if possible)
- Try to make it interesting
- Explain why the new knowledge is valuable

Popular Scientific Writing Advice

- Structure
  - Describe *why* and *how* the research will be conducted, and explain *in what way* the new knowledge might be important
- Tone
  - Avoid using overly informal speech; Magazine style not social media

    Know your audience
  - If you want to write in a more conversational style, you have to be much more careful
- Proofreading
  - In a short abstract, clarity is everything
  - Small mistakes are at best annoying, at worst confusing

Technical vocabulary

- Linguistic vocabulary
  - ”Morphology”, ”Corpus”, ”Natural Language”
- Computational vocabulary
  - ”Model”, ”Machine Learning”, ”Python module”
- Mathematical vocabulary
  - ”Vector in Vector-space”

Technical vocabulary

- Linguistic vocabulary
  - "Morphology", "Corpus", "Natural Language"
- Computational vocabulary
  - "Model", "Machine Learning", "Python module"
- Mathematical vocabulary
  - "Vector in Vector-space"
- Some strategies
  - Make sure that key elements are introduced gradually
  - Use a very simply phrased definition or a restatement in simple terms, introduced as close to the term as possible
  - Use examples

Implementing the Project

- Start up: mobilize project resources
  - Hire researchers and other personnel
  - Acquire equipment, software, data, literature
- Manage research activities
  - Implement project plan
  - Revise plans if necessary
- Close down: report project outcome
  - Dissemination of results (publications)
  - Report to funding agency

Your Project Proposals

- Maximum 3 pages excluding references (and not much shorter)
- Structure (following the VR Guidlines):
  - Purpose and aims (max 0.5 page)
  - Survey of the field (max 1 page)
  - Project description (min 1.5 page)
    - Theories and methods
    - Time plan and implementation
  - References
- Use the LaTeX template available!

Your Projects

- Time to start thinking seriously about a project
  - Proposals due **October 6**
    - Main proposal: 3 pages plus references (pdf)
    - Popular science abstract: 2000 characters (txt)
  - Presentations October 11 (8 minutes with slides)
    - Email your slides as pdf-files to your group leader the day before the seminar!
- Contact your group leaders if you need advice

Project Advice

- Data, data, data
  - It is much easier if you can find a ready to use data set
  - Annotation is dificult, and time consuming
  - We have had annotation projects in the past, but we do not reommend them
  - A small annotation effort is possible, but try to avoid it

Project Advice

- Data, data, data
  - It is much easier if you can find a ready to use data set
  - Annotation is dificult, and time consuming
  - We have had annotation projects in the past, but we do not reommend them
  - A small annotation effort is possible, but try to avoid it
  - Look at shared tasks
  - Example: SemEval

Project Advice

- Data, data, data
  - It is much easier if you can find a ready to use data set
  - Annotation is dificult, and time consuming
  - We have had annotation projects in the past, but we do not reommend them
  - A small annotation effort is possible, but try to avoid it
  - Look at shared tasks
  - Example: SemEval
- Think about evaluation

Project Advice

- Data, data, data
  - It is much easier if you can find a ready to use data set
  - Annotation is dificult, and time consuming
  - We have had annotation projects in the past, but we do not reommend them
  - A small annotation effort is possible, but try to avoid it
  - Look at shared tasks
  - Example: SemEval
- Think about evaluation
- Research questions!
  - Formulate them early
  - Make sure that your proposed experiments can answer them
  - You may modify your research questions during your project

Computational Resources

- You can use the Snowy cluster at UPPMAX if you need extra computational resources
  - Course project:
    - UPPMAX 2022/2-17
    - Gives you extra storage
      Create a personal folder under: `/proj/uppmax2022-2-17`
  - General CL project:
    - UPPMAX 2020/2-2
    - Gives you priority in the GPU queue
- Create an account at `supr.snic.se` and apply for the projects
- Read up on using UPPMAX and the Slurm queueing system!

Coming up

- Next week
  - Final literature seminar
    - In case you missed a seminar you should have gotten instructions for how to compensate
- Project proposal

Exercise: re-write as a popular science abstract

We show how we can adapt a dependency parser to low-resource domains by combining treebanks across languages. We demonstrate how we can take advantage of in-domain treebanks from other languages, and show that this is especially useful when only out-of-domain treebanks are available for the target language. Two parameter-free methods for applying treebank embeddings at test time are proposed, which give competitive results to tuned methods when applied to Twitter data and transcribed speech. This gives us a method for selecting treebanks and training a parser targeted at any combination of domain and language.