NNT : 2017SACLS027

# Thèse de doctorat de l'Université Paris-Saclay préparée à l'Université Paris-Sud

Ecole doctorale n°580
Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat: Informatique

par

## Mme Dialekti Valsamou

Extraction d'Information pour les réseaux de régulation de la graine chez *Arabidopsis Thaliana.*

Thèse présentée et soutenue à Orsay, France, le 17 janvier 2017.

Composition du Jury :

| | | | |
|---|---|---|---|
| M. | Bertrand Dubreucq | Directeur de Recheche<br>INRA, Institut Jean-Pierre Bourgin | (Président du jury) |
| Mme. | Pascale Sébillot | Professeur<br>INSA de Rennes | (Rapporteur) |
| Mme. | Isabelle Tellier | Professeur<br>Université Paris 3 | (Rapporteur) |
| M. | Jean-Philippe Vert | Directeur de Recherche<br>Mines ParisTech | (Examinateur) |
| M. | Pierre Zweigenbaum | Professeur<br>Université Paris-Sud | (Co-directeur de thèse) |
| Mme. | Claire Nédellec | Directeur de Recherches<br>INRA Jouy-en-Josas | (Co-directrice de thèse) |

# Acknowledgments

I would like to start this acknowledgment section by expressing my deepest gratitude to my two advisors, Claire Nédellec and Pierre Zweigenbaum. Their continuous support and guidance throughout this work have been instrumental.

Special thanks are due to Bertrand Dubreucq, Loïc Lépiniec and Philippe Bessières for their help and guidance on all things *Arabidopsis Thaliana* and biology, as well as their encouragement.

I would equally like to thank my colleagues in both MaIAGE and LIMSI labs, and namely Robert Bossy, Estelle Chaix, Zorana Ratkovic, Abdou Fatihi, Louise Deléger, Wiktoria Golik, Pierre Warnier and Julien Jourde, for both our collaboration and friendship.

Last but not least, I want to thank my family and friends for their love, support and encouragement, without which this task would have impossible.

*to Alexandre.*

*You make me better.*

# Contents

**List of Figures** **xi**

**List of Tables** **xiii**

**Abstract** **1**

**Introduction** **3**

1 Problem Statement . . . . . . . . . . 3

1.1 Background of the Problem . . . . . . . . . . 3

1.2 Purpose and Significance of the Study . . . . . . . . . . 4

2 Research Questions . . . . . . . . . . 5

3 Thesis Contributions . . . . . . . . . . 5

3.1 Research Design and Limitations . . . . . . . . . . 6

4 Thesis Outline . . . . . . . . . . 6

**1 Background & Related Work** **9**

1.1 Introduction . . . . . . . . . . 10

1.2 Biological Background . . . . . . . . . . 10

1.2.1 Why the seed development of Arabidopsis thaliana? . . . . . . . . . . 10

1.2.2 *A. thaliana* regulatory network basics . . . . . . . . . . 11

1.3 The seed development network in *A. Thaliana* . . . . . . . . . . 12

1.4 Knowledge Extraction . . . . . . . . . . 14

1.4.1 Knowledge models, domains and IE . . . . . . . . . . 14

1.4.2 Knowledge Extraction for Biology and *A.Thaliana* . . . . . . . . . . 17

1.5 Knowledge Expressed in Text: the Corpus . . . . . . . . . . 19

1.5.1 The building blocks: entities and semantic relations . . . . . . . . . . 19

1.5.2 Corpus . . . . . . . . . . 20

1.6 Information Extraction . . . . . . . . . . 25

1.6.1 Defining IE . . . . . . . . . . 25

1.6.2 Evaluating IE systems . . . 26
1.6.3 IE Systems . . . 27
1.7 The Alvis ecosystem . . . 34
1.8 Conclusion . . . 35

**2 Data** **37**
2.1 Introduction . . . 38
2.1.1 Tasks & Development Phases . . . 38
2.1.2 People . . . 39
2.2 Tools . . . 40
2.2.1 Collaborative Tools . . . 40
2.2.2 Documentation . . . 41
2.2.3 Annotation . . . 43
2.3 Model . . . 43
2.3.1 Conceptual Model . . . 48
2.3.2 Annotation Model . . . 55
2.3.3 Model Transformations . . . 60
2.4 Corpus . . . 64
2.4.1 Source . . . 64
2.4.2 Corpus Annotation . . . 68
2.5 Results and Discussion . . . 70
2.6 Conclusion . . . 74

**3 Relation Extraction** **77**
3.1 Introduction . . . 78
3.1.1 Information Extraction . . . 79
3.2 Linguistic Pre-processing . . . 81
3.2.1 Tokenization and segmentation . . . 81
3.2.2 Lemmatization and normalization . . . 81
3.2.3 Syntactic Parsing . . . 81
3.3 Representation . . . 82
3.3.1 Introduction . . . 82
3.3.2 From text files to complex sentence objects . . . 82
3.3.3 From sentences to candidate relations . . . 84
3.3.4 From candidate relations to paths . . . 85
3.3.5 From paths to a machine-learning ready representation . . . 91
3.4 Classification . . . 98

3.4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
3.4.2 Support Vector Machines . . . . . . . . . . . . . . . . . . . . . . 99
3.5 Adding Semantic Information . . . . . . . . . . . . . . . . . . . . . . 101
3.5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 101
3.5.2 Manual Classes . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
3.5.3 WordNet . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
3.5.4 Textual similarity . . . . . . . . . . . . . . . . . . . . . . . . . . 103
3.5.5 Distributional Semantics . . . . . . . . . . . . . . . . . . . . . . 103
3.6 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
3.6.1 Experimental setup . . . . . . . . . . . . . . . . . . . . . . . . . 106
3.6.2 Experimental validation . . . . . . . . . . . . . . . . . . . . . . . 106
3.7 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116
3.8 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 117

**4 Relation Extraction on *A. thaliana*** **119**
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 119
4.2 SeeDev in BioNLP-ST '16 . . . . . . . . . . . . . . . . . . . . . . . . 120
4.3 AlvisRE on *Arabidopsis thaliana* . . . . . . . . . . . . . . . . . . . 122
4.3.1 Optimizing the SVM margin parameter . . . . . . . . . . . . . . 122
4.3.2 Exploring different relation types . . . . . . . . . . . . . . . . . 122
4.3.3 The impact of model transformations on Relation Extraction . . . . . . . 124
4.3.4 AlvisRE compared to the SeeDev participants . . . . . . . . . . . 125
4.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126

**Conclusion & Future Work** **131**
1 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 131
2 Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133
2.1 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133
2.2 Information Extraction . . . . . . . . . . . . . . . . . . . . . . . . . 133

**Appendices** **137**

**A Manual Classes for LLL** **139**

**B Trigger words for Bacteria Biotopes** **141**

**C A complete AlvisRE input file** **143**

**D The list of the articles in the *Arabidopsis Thaliana* corpus.** **157**

**E The *Arabidopsis Thaliana* Seed Development conceptual model in detail. 161**

**F The *Arabidopsis Thaliana* Seed Development annotation model in detail. 177**

**Bibliography 183**

# List of Figures

1.1 *Arabidopsis thaliana* plant in the Jardin de Plantes in Paris & *A. thaliana* flower. 11
1.2 A regulatory network in *Bacilus Subtilis* . . . . . . . . . . . . . . . . . . . . . . . . . 13
1.3 Regulatory steps in *Arabidopsis* seed development . . . . . . . . . . . . . . . . . . 14
1.4 From Data to Information to Knowledge to Wisdom. . . . . . . . . . . . . . . . . . 15
1.5 Complexity in knowledge modelization. . . . . . . . . . . . . . . . . . . . . . . . . . 16
1.6 PPI Example . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
1.7 Alvis Suite Architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

2.1 Redmine . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
2.2 Annotation Wiki . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
2.3 AlvisAE: annotation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44
2.4 AlvisAE: campaign & workflow . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45
2.5 AlvisAE: consolidation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46
2.6 Adjudication tool . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46
2.7 Entity and relation adjudication . . . . . . . . . . . . . . . . . . . . . . . . . . . . 47
2.8 Conceptual Entity type hierarchy . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51
2.9 Annotation Entity type hierarchy . . . . . . . . . . . . . . . . . . . . . . . . . . . . 57
2.10 Relation type schema . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 58
2.11 Constraint table for “Regulates Activity Of” . . . . . . . . . . . . . . . . . . . . . . 61
2.12 Splitting “Regulates Activity Of” . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
2.13 Transformation of a condition relation. . . . . . . . . . . . . . . . . . . . . . . . . . 63
2.14 Corpus Sources . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65
2.15 Yearly distribution of the document packs. . . . . . . . . . . . . . . . . . . . . . . . 66
2.16 Corpus document length . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 66
2.17 Distribution of passages in typical sections . . . . . . . . . . . . . . . . . . . . . . . 67
2.18 Corpus passage length . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
2.19 Inter-annotator agreement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 70
2.20 Manually Annotated Entities . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
2.21 Manually Annotated Relations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
2.22 Relations in Seedev . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74

3.1 The file format . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83
3.2 Sentence representation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85
3.3 Paths from the Bacteria Biotopes Corpus . . . . . . . . . . . . . . . . . . . . . . . 88
3.4 Representation: "Entire Sentence" . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
3.5 Representation: "Surface Path" . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
3.6 Fixing parsing problems . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90

3.7 A path rewriting. . . . . 92
3.8 Gap penalty . . . . . 93
3.9 Alignment of two candidates . . . . . 95
3.10 The EKM transformation . . . . . 97
3.11 Linear and non-linear separators for classification. . . . . . 100
3.12 A non-linear input space transformed by a kernel function . . . . . 100
3.13 Support Vectors . . . . . 100
3.14 The data model of AlvisNLP/ML . . . . . 107
3.15 The impact of the number of candidates . . . . . 109

# List of Tables

2.1 Inter-annotator agreement . . . . . . . . . . . . . . . . . . . . . . . . . . . . 69
2.2 Seedev corpus statistics . . . . . . . . . . . . . . . . . . . . . . . . . . . . 71
2.3 Relation types in the Seedev Corpus . . . . . . . . . . . . . . . . . . . . . . 72

3.2 Results on the LLL corpus . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
3.3 Results on the BB Task 2 corpus . . . . . . . . . . . . . . . . . . . . . . . . 111
3.4 BioNLP'13 BB Task 2 participant evaluation. . . . . . . . . . . . . . . . . . . 112
3.5 A study on LLL using AlvisRE with Global Alignment and semantic similarities. 114
3.6 Using manual classes and word embeddings on LLL . . . . . . . . . . . . . . . 114
3.7 Results on BB Task 3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 115
3.8 BioNLP'13 BB Task 3: Official scores vs AlvisRE . . . . . . . . . . . . . . . . 116

4.1 Official Results on SeeDev-binary: by participant . . . . . . . . . . . . . . . . 121
4.2 Official Results on SeeDev-binary: by relation type . . . . . . . . . . . . . . . 121
4.3 Cross Validation on seedev-binary . . . . . . . . . . . . . . . . . . . . . . . . 123
4.4 The impact of the test dataset. . . . . . . . . . . . . . . . . . . . . . . . . . 124
4.5 Results using the annotation model . . . . . . . . . . . . . . . . . . . . . . . 125
4.6 Results on seedev-binary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 127
4.7 Results on seedev-binary: word2vec . . . . . . . . . . . . . . . . . . . . . . . 128

# Abstract

While information is abundant in the world, structured, ready-to-use information is rare. This work proposes Information Extraction (IE) as an efficient approach for producing structured, usable information on biology, by presenting a complete IE task on a model biological organism, *Arabidopsis thaliana*. Information Extraction is the process of extracting meaningful parts of text and identifying their semantic relations.

In collaboration with experts on the plant *A. Thaliana*, a knowledge model was conceived. The goal of this model is providing a formal representation of the knowledge that is necessary to sufficiently describe the domain of grain development. This model contains all the entities and the relations between them which are essential and it can directly be used by algorithms. In parallel, this model was tested and applied on a set of scientific articles of the domain. These documents constitute the corpus which is needed to train machine learning algorithms. The experts annotated the text using the entities and relations of the model. This corpus and this model are the first available for grain development and among very few on *A. Thaliana*, despite the latter's importance in biology. This model manages to answer both needs of being complex enough to describe the domain well, and of having enough generalization for machine learning.

A relation extraction approach (AlvisRE) was also elaborated and developed. After entity recognition, the relation extractor tries to detect the cases where the text mentions that two entities are in a relation, and identify precisely to which type of the model these relations belong to. AlvisRE's approach is based on textual similarity and it uses all types of information available: lexical, syntactic and semantic. In the tests conducted, AlvisRE had results that are equivalent or sometimes better than the state of the art. Additionally, AlvisRE has the advantage of being modular and adaptive by using semantic information that was produced automatically. This last feature allows me to expect similar performance in other domains.

**Keywords:** Information Extraction, Relation Extraction, Natural Language Processing, NLP, BioNLP, Bioinformatics.

# Introduction

**1 Problem Statement** . . . . . . . . . . . . . . . . . . . . . . . . . . . . **3**

1.1 Background of the Problem . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2 Purpose and Significance of the Study . . . . . . . . . . . . . . . . . . . . 4

**2 Research Questions** . . . . . . . . . . . . . . . . . . . . . . . . . . . . **5**

**3 Thesis Contributions** . . . . . . . . . . . . . . . . . . . . . . . . . . . **5**

3.1 Research Design and Limitations . . . . . . . . . . . . . . . . . . . . . . . 6

**4 Thesis Outline** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **6**

## 1 Problem Statement

Readily available, structured knowledge in the biological domain is limited and unsatisfactory for the experts. The limitations of traditional approaches of manual curation and experimental data collection make scientific text a good alternative as a source of information.

Information Extraction (IE) has been tested previously on biological text with positive results, but as it starts being further explored as an alternative, the needs of the biologists call for more complex model and data in order to adequately represent domain knowledge.

This work presents a complete IE task on a model biological organism, *Arabidopsis thaliana*, in order to show how Information Extraction can be a good approach for producing structured, usable information on biology. It does so by introducing a knowledge model and an annotated corpus that correspond well to the needs of experts in the *Arabidopsis thaliana* domain, along with an algorithmic approach that is adapted to this task.

### 1.1 Background of the Problem

Giving data meaning by establishing relational connections transforms raw, meaningless data into usable information. By adding context and structure this information is, in turn, transformed into knowledge. IE is the process of extracting structured information from textual data in natural language. While these texts contain knowledge in the eyes of the human reader, they are raw, unusable data for the computer. IE can translate these raw data into information fast and make the human reader's work much easier by providing an appropriate structured representation. Having structured information is not just helpful for direct use by the biologist, but also it is a necessary for any formal representation required by bioinformatics applications.

The limitations of non text-based approaches come from the fact that these use experimental and manually curated data in order to produce models and databases, procedures which are both costly and time consuming. The interpretation of experimental results and the maintaining curated databases in the long term are both non-trivial tasks. Scientific literature, on the other hand, has the advantage of containing a lot of detailed, up to date, and contextualized information and presents results along with their interpretation. This contrast in the availability of structured knowledge bases versus scientific texts on biology makes IE particularly interesting for biologists. Inversely, the domain has the advantages of well-established norms in scientific writing, rich resources like ontologies and an awareness of the benefits of IE in the community, thus making biology an attractive domain of application for IE researchers. Text-based approaches can be used independently, but they can also be complementary to other sources for information.

*Arabidopsis thaliana* is a model plant of great value for both the scientific community and the agricultural industry. As a model organism, it has been and continues to be heavily studied, and research done on *A. thaliana* has the possibility of being generalized to other plants. The plant and its seeds, in particular, are equally very important for the industry, as they constitute a major source of human and animal nutrition.

Finally, in spite of the fact that IE on biology has been gathering a lot of interest in the last two decades, the benchmarks used for evaluation have mostly been less complex than a problem focused on the needs of the end-users –the biologists, such as the one tackled by this work. The international challenges organised on the domain started by proposing simpler entity and relation extraction tasks, such as the detection of protein mentions and their interactions, and have steadily been advancing towards the extraction of more intricate structures such as biological networks. The domain is currently reaching a level of maturity that allows it to envision and implement applications that correspond to the needs of biologists. More demanding models call for more adapted algorithmic solutions, that look into the future, improve results and, at the same time, are designed and developed with the biologists' needs in mind.

This study focuses on the Relation Extraction (RE) subtask of IE. This is mainly due to the fact that the aforementioned increase in complexity of bio-IE tasks concerns mainly relations. Consequently, it was deemed more fitting to focus on RE and draw on the work of colleagues working on closely related topics for the other subtasks of IE.

## 1.2 Purpose and Significance of the Study

The purpose of this work is to successfully perform the complete task of IE on *A. thaliana*. This includes producing the model, the data and the algorithm necessary for IE on *A. thaliana*. Its significance comes from the fact that this work was done in close collaboration with biologists on the model plant, *A. thaliana*. This collaboration insured that this work remains significant for its domain of application, biology. At the same time, no compromises were made with regards to the quality of IE and therefore the improvements in Relation Extraction proposed here are of interest to the IE community, regardless of the domain of application.

## 2 Research Questions

This thesis was focused on two primary research questions:

1. Can there be a compromise in data and model complexity in order to both adequately describe a biological domain of knowledge, and still be able to use machine learning approaches?
2. How can we maximize the use of different types of knowledge within the algorithmic approach in order to generalize best and increase performance in complex tasks, but also be easily adaptable to new tasks?

Regarding these two primary axes, the hypotheses of this thesis are that:

1. By employing an iterative process and involving a multidisciplinary team of experts, one can attain both goals and produce high quality re-usable models and data with a level of detail that satisfies the needs of the biologist and yet insure that the volume and generalization of the model and the data is adapted to machine learning. And,
2. By using different types of information in a modular way, a Relation Extraction method can perform well in different, complex tasks.

## 3 Thesis Contributions

I make the following contributions in this dissertation:

- I contribute to the introduction of a new corpus for IE on a model organism.
    - This corpus is built with an application in mind, thus making it closer to what an actual real-world IE task would use.
    - It was built in close collaboration with biologists
    - This corpus is focused on a model organism, making generalization to other plants a possibility.
    - In particular, *A. thaliana* is a plant that is important for both scientific community and industry.
- The model and data are published in an international challenge and publicly available for reuse and evaluation.
- I examine and evaluate different levels of information in RE: lexical, syntactic and semantic, using external tools.
    - In particular, I use recent developments in unsupervised methods for producing semantic information like word2vec, enabling the use of domain-adapted semantic knowledge.
- I examine and evaluate the effects of imperfect syntactic analysis tools and propose and evaluate a solution to this problem.

- I propose a ML-based RE method that is modular and can be tuned to each problem but also incorporate other tools and methods.
- I produce a RE method which outperforms current state of the art.
- The RE method is designed to work within the Alvis pipeline, further enriching the latter.

### 3.1 Research Design and Limitations

*Procedures & Research Design*

Three principal components can be identified for this work: the knowledge model, the corresponding textual corpus and the IE method. These three components correspond to three separate, intermediate goals.

The first goal is model design. This was, inevitably, the starting point of this work and it was done in close collaboration with the biologists. After an initial model version which was developed in theory, repeated iterations of application of each model instance to data annotation allowed us to reach a convergence in the design of the model.

The second goal was the annotation of the textual corpus. After constituting a set of selected scientific publications, the experts started annotating an initial subset of this text using the model. Once the model was finalised, the rest of the corpus was annotated. Additionally, some transformations were introduced in order to adapt model and data to different configurations.

In parallel with the iterative development of the first two goals, I worked on the RE algorithm. For the evaluation of this algorithmic approach I used existing corpora from past international challenges on IE for biology.

Once these three intermediate goals were accomplished, the proposed algorithmic approach was used on the newly created *A. thaliana* corpus. I was fortunate enough to have this corpus included in a recent international challenge, thus enabling me to compare my results to the state of the art.

*Assumptions and Limitations*

An inherent assumption of this thesis is the authority of the collaborating experts on the *A. thaliana* domain. Additionally, this thesis assumes that the international challenges used for evaluation and comparison represent the state of the art on IE for biology.

As a final note, while this allowed the RE algorithm to be tested on other data, an inevitable limitation of this thesis was that the target data were being produced in the course of this work, so I was only able to combine algorithm and data towards the end of my work on this thesis.

## 4 Thesis Outline

The remainder of the dissertation is organized in the following way:

- *Chapter 1* covers the biology and information extraction **Background** that is necessary to put this work in context. Consequently, it presents the **Related Literature**, as well as any external tools used.
- *Chapter 2* is dedicated to **Data.** It presents the design process for the model and the annotation of data, as well as their transformations. Additionally, it provides information and statistics about the produced corpora.
- *Chapter 3* details the **Relation Extraction** method developed and documents design decisions by giving intermediary results on established datasets.
- *Chapter 4* presents the **Results** of the algorithmic approach on the data on *A. thaliana.*
- Finally, **Conclusions** and an insight on **Future Work** can be found at the end of this thesis.

# Chapter 1

# Background & Related Work

**1.1 Introduction** . . . . . . . . . . **10**

**1.2 Biological Background** . . . . . . . . . . **10**

1.2.1 Why the seed development of Arabidopsis thaliana? . . . . . . . . . . 10

*A. thaliana* . . . . . . . . . . 10

Seeds . . . . . . . . . . 11

1.2.2 *A. thaliana* regulatory network basics . . . . . . . . . . 11

What is a regulatory network . . . . . . . . . . 11

*A. thaliana* and regulatory networks. . . . . . . . . . . 12

**1.3 The seed development network in *A. Thaliana*** . . . . . . . . . . **12**

**1.4 Knowledge Extraction** . . . . . . . . . . **14**

1.4.1 Knowledge models, domains and IE . . . . . . . . . . 14

1.4.2 Knowledge Extraction for Biology and *A.Thaliana* . . . . . . . . . . 17

Importance . . . . . . . . . . 17

A short historical survey of projects and applications . . . . . . . . . . 19

**1.5 Knowledge Expressed in Text: the Corpus** . . . . . . . . . . **19**

1.5.1 The building blocks: entities and semantic relations . . . . . . . . . . 19

1.5.2 Corpus . . . . . . . . . . 20

Corpus Design *&* Annotation . . . . . . . . . . 21

General Purpose Corpora . . . . . . . . . . 22

Corpora Specialized in Biology . . . . . . . . . . 23

**1.6 Information Extraction** . . . . . . . . . . **25**

1.6.1 Defining IE . . . . . . . . . . 25

What IE is . . . . . . . . . . 25

What IE is not . . . . . . . . . . 26

1.6.2 Evaluating IE systems . . . . . . . . . . 26

1.6.3 IE Systems . . . . . . . . . . 27

Overview: Two axes of evolution . . . . . . . . . . 27

Named Entity Recognition . . . . . . . . . . . . . . . . . . . . . . . . . . 28
Coreference Resolution . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
Relation Extraction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
**1.7 The Alvis ecosystem . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34**
**1.8 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35**

## 1.1 Introduction

*"The primary goal of text mining is to retrieve knowledge that is hidden in text and to present the distilled knowledge to users in a concise form."*

—Hearst [Hearst, 1999].

Scientific literature offers a rich source of highly specialized knowledge. By extracting information from scientific articles we have access to the synthesis, analysis and interpretation of experimental data. To accomplish this task, it is necessary to create representative knowledge models, and adapted methods of extraction. The models and methods used need at the same time to be domain specific and able to generalize.

Information extraction from scientific literature for the elaboration of biological networks is an interdisciplinary task. This chapter will cover the necessary background and related literature for the different domains involved.

The chapter first introduces the biological context of this work followed by the three principal components and goals of this work: the knowledge model, the corresponding textual corpus and the IE method.

## 1.2 Biological Background

### 1.2.1 Why the seed development of Arabidopsis thaliana?

#### *A. thaliana*

*Arabidopsis thaliana* is small flowering plant that grows in Eurasia, and even though it has little economic importance, it is a model organism that has been extensively studied, as it combines a great number of desirable traits.

With one of the smallest genomes in plants and various practical advantages, *A. thaliana* was the first plant to have its genome fully sequenced, and extensive genetic and physical maps produced for its 5 chromosomes. It has a very rapid life cycle and it requires very little space and nutrients to grow. The extent to which its genome has been studied and the resources available make genetic engineering with *A. thaliana* easier and faster than any other plant [TAIR, 2015, NSF, 2013]. The physical map of all chromosomes has been completed in 1997 and the complete genome sequencing has been achieved in 2000.

Figure 1.1: *Arabidopsis thaliana* plant in the Jardin de Plantes in Paris & *A. thaliana* flower.

Studying *Arabidopsis thaliana,* the model plant, allows researchers to gain comprehensive knowledge of a complete plant and enrich an established reference system. Additionally, *Arabidopsis thaliana* is considered to be so similar to the majority of other plants that any discovery made on it us likely to generalize to all flowering plants [NSF, 2000].

#### Seeds

Seeds play a fundamental role in agriculture. They are the vector for breeding and production of crops, but they are also the part of the plant that is most used for human and animal consumption and industrial uses such as energy from plant oils [Baud et al., 2002]. Consequently, knowledge in seed biology is of great importance for the scientific community and the industry [North et al., 2010].

Seeds constitute the evolutionary advantage of spermatophytes (seeding plants), allowing them to interrupt and resume their life cycles depending on the environmental conditions [Bentsink et al., 2008, Bewley, 1997]. In the case of *Arabidopsis,* the abundance of resources has allowed the scientific community to develop molecular genetic approaches on seeds and study in depth the gene regulatory networks that control seed development and maturation [North et al., 2010].

### 1.2.2 *A. thaliana* regulatory network basics

#### What is a regulatory network

Biological networks achieve biological functions, they are often represented as graphs. Even though there exist some types of non-cellular networks, intra-cellular networks are much more frequent. The goal of such networks is to describe the complex relationships among biological agents taking place within a cell that regulate the behaviour of cells, organs or even organisms by extension [Blais and Dynlacht, 2005, Barabasi and Oltvai, 2004].

The four most common types of biological networks are (i) gene regulatory networks (GRNs), (ii) protein-protein interaction (PPI) networks, (iii) signaling (iv) metabolic pathway networks. Respectively, they describe (i) the activation or inhibition relationships between genes, (ii) the physical interactions between proteins as well as the metabolic and signaling pathways of the cell, (iii) cell communications by extracellular signaling and cell responses and (iv) the metabolic products and substrates that participate in one reaction. Finally, hybrid networks containing integrated information of gene regulation, signaling and metabolic pathways can also be found [Lee et al., 2008].

Historically, GRNs have been the focus of scientific research. They were first proposed by [Monod and Jacob, 1961] as a concept and have evolved from purely theoretical models to robust tools being used regularly in both theoretical and experimental studies [Kauffman, 1969, Espinosa-Soto et al., 2004, Aldana and Cluzel, 2003].

In such network representations the nodes correspond to genes, proteins, or more generally any biomolecule. Edges can represent direct molecular interactions such as bindings of proteins of gene promoters, regulatory interactions, or the sharing of functional properties. They can be very complex even when they are highly localized, as the various types of interactions can have conditions of varied nature (quantitative, environmental, etc). Figure 1.2 shows the complexity of one such network in *Bacillus subtilis* bacteria [Goelzer et al., 2008].

The complexity of networks involving multiple genes and environmental factors is such that the understanding of such mechanisms are still an open question in biology [Alvarez-Buylla et al., 2007]. The ultimate goal of systems biology is precisely the inference of the behaviour of regulatory networks in new conditions [Krouk et al., 2013]. This is even more true in the case of plants, whose rapid adaptation to environmental changes and subsequent changes in gene expression result in very complex networks [Krouk et al., 2013].

#### ***A. thaliana* and regulatory networks.**

The regulatory networks involved in a number of functions and development phases of *Arabidopsis Thaliana* have been studied and a great number of articles describing these regulations have been published (examples include: [Mendoza et al., 1999, Espinosa-Soto et al., 2004, Balbi and Devoto, 2007]).

## 1.3 The seed development network in *A. Thaliana*

The molecular and genetic mechanisms involved in seed development are elaborate, necessitating the coordination of different genetic materials, the development of multiple tissues and a number of environmental interactions. A schematic representation of these regulatory steps can be found in Figure 1.3 [North et al., 2010]. Currently, our understanding of this regulatory network involved in seed development in *Arabidopsis thaliana* is far from complete [Santos-Mendoza et al., 2008], even if it constitutes a very active research domain [Baud et al., 2002, Weijers and Jürgens, 2005, Lepiniec et al., 2006, Dubreucq et al., 2009].

Existing research describes a complex regulatory network with a number of identified master regulatory genes. While the function of some of these regulators remains unknown (*e.g* TAN and L1L), four genes have been identified as having control over seed maturation: LEC1 and LEC2, FUS3 and ABI3. These regulators do not act independently, but are part of an intricate scheme together with hormones, epigenetic mechanisms and target regulatory proteins. Interestingly, these master regulators can control seed development directly by the accumulation

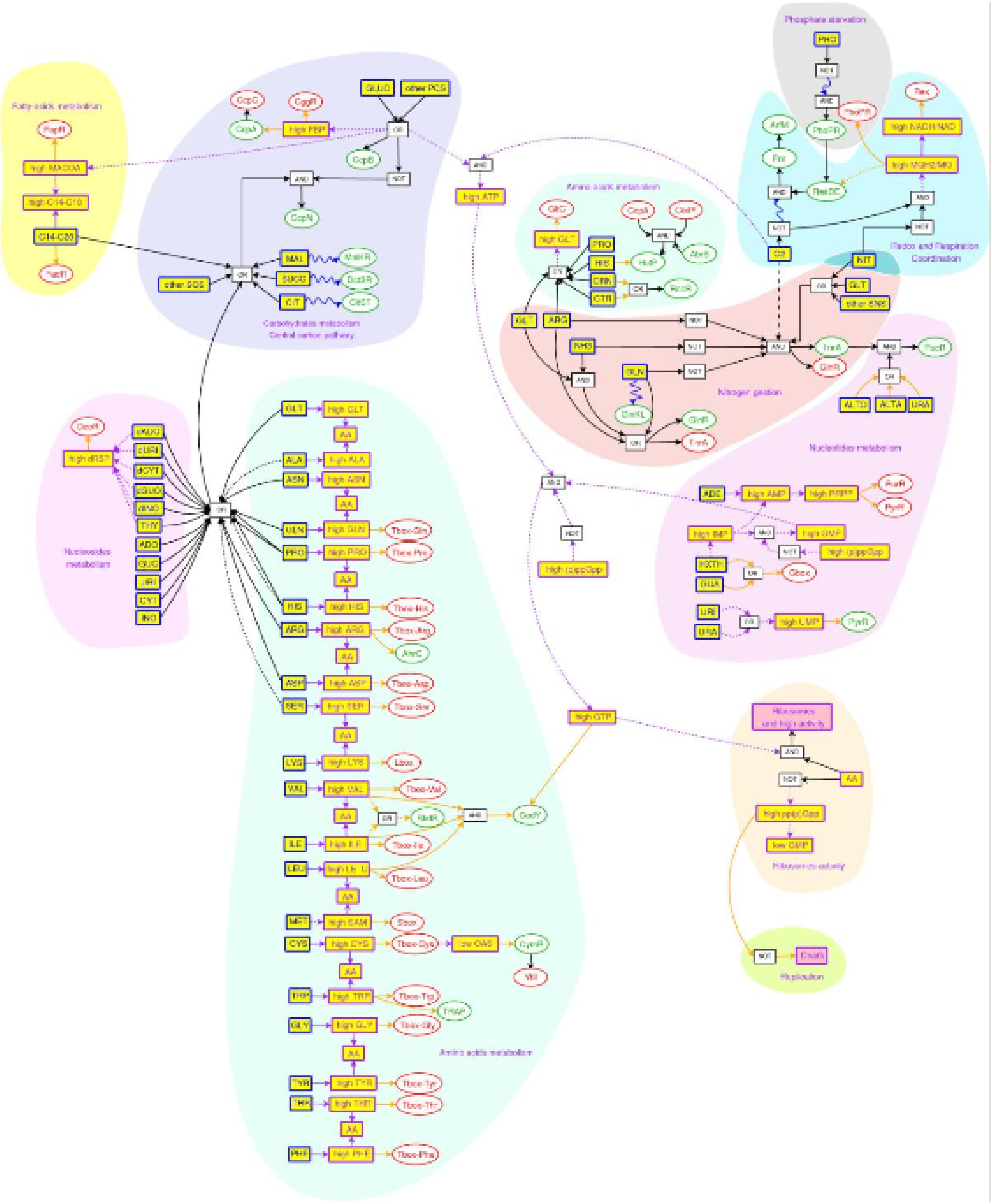


Figure 1.2: Visualization of a regulatory network in *Bacillus Subtilis* [Goelzer et al., 2008].

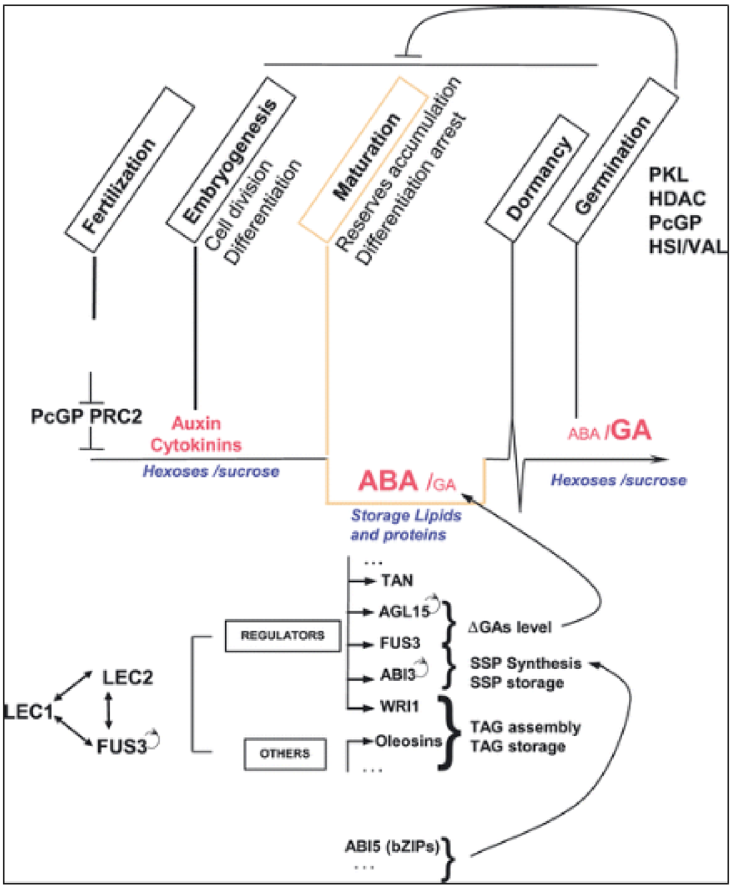


Figure 1.3: Schematic representation of the regulatory steps controlling *Arabidopsis* seed development.

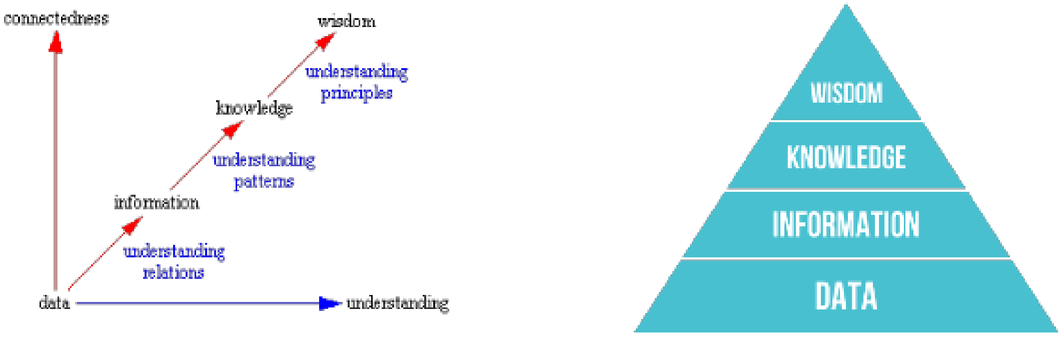
of biomolecules, and/or indirectly by the activation of secondary transcription factors able to trigger other transcription programs. One can hypothesize that this complex network provides a robust and tight control of seed maturation [Dubreucq et al., 2009].

The importance of Arabidopsis for both science and industry and the role of the seed, in particular, make this regulatory network an ideal candidate for further research. The complexity of the network and the partial knowledge we currently have on it call for an integrative approach which can bring together the various existing sources of information and aim at giving a global view of the shared knowledge on the topic.

## 1.4 Knowledge Extraction

### 1.4.1 Knowledge models, domains and IE

**Data, Information, Knowledge,** and **Wisdom** are the transformation steps that take us from raw facts or signals to understanding. The DIKW pyramid (Fig. 1.4b) has often been used to depict the relationships between them. Bellinger et al.[Bellinger et al., 2004] expanded on the definitions proposed by Ackoff [Ackoff, 1989] and proposed the diagram in Figure 1.4a to explain the transformation. Whereas wisdom falls outside the scope of knowledge extraction, the other



(a) [Bellinger et al., 2004]

(b) The DIKW pyramid

Figure 1.4: From Data to Information to Knowledge to Wisdom.

three concepts are fundamental notions for the domain.

Raw **data** simply exists, devoid of any significance. **Information** is data that has been given meaning by relational connections. **Knowledge**, finally, requires context, organization and structure and even though it is generally considered hard to define [Rowley and Hartley, 2008], it is often tied to a notion of application, in the sense that knowledge is intended to be useful for a given task.

In order to illustrate these nuances in the case of knowledge extraction from text, consider the following sentence: *"LEC1 and LEC2 are specifically expressed in seeds"*. Starting with the raw data seen as strings of characters, and following their transformation, even after having detected the individual words and the occurrence of genes LEC1 and LEC2 in this phrase, it is still considered data before detecting any relation. Once a human reader or a computer program has understood that there exists a relation of expression between these genes and seeds, we can talk about information. But it is only when this information is put in context by understanding through experience, or an appropriate knowledge model, that we can consider this knowledge.

A **knowledge model** is a formal, consistent representation of knowledge. It can be described using logic, tabular representations, a diagram or graph or any other structured representation of concepts or pieces of knowledge and the relationships between them with a formal semantics attached. Knowledge models have played an important role for decades in the field of Artificial Intelligence and they have been used for knowledge acquisition and engineering applications, decision support, expert systems and a number of other tasks.

The purpose of a knowledge model is to adequately represent the knowledge of the domain or subdomain it describes, and at the same time to provide a representation allowing reasoning and simulation. Explanatory and predictive models of knowledge allow scientists to summarize and explain, share knowledge, formally verify hypotheses and formulate new ones. Knowledge models are generally task-oriented because the representation choices must be driven by the future use.

Various types of models exist, each serving different purposes and necessitating different levels of detail, hierarchy and formality. In Figure 1.5, some typical examples of models are listed by order of complexity and logical formalism. More formal models allow us to calculate the truth value of an assertion, to derive new rules and facts and guarantee formal properties, such as consistency, completeness and minimality.

In knowledge models, knowledge is generally represented as concepts, groupings, relations

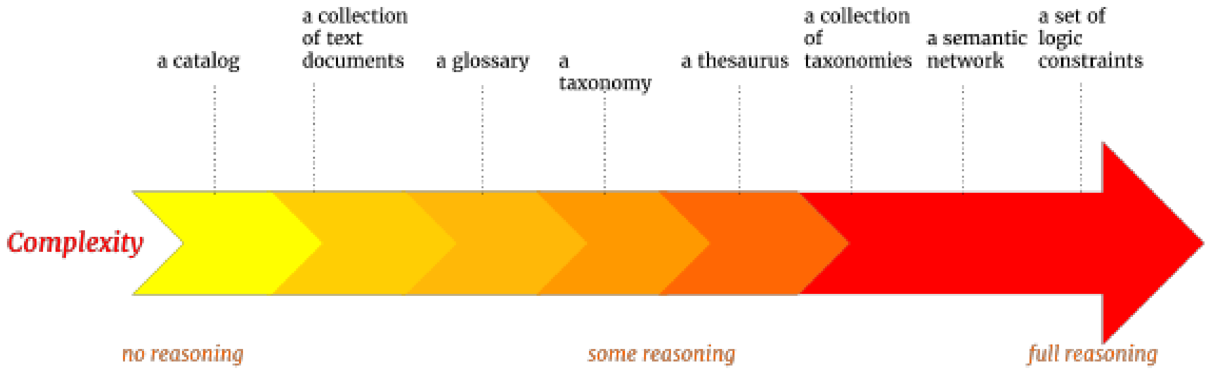


Figure 1.5: Complexity in knowledge modelization.

between concepts and, optionally, rules and instances. Concepts and their relations define types of information and the valid relationships between them. They are the abstraction layer which provides the structure and organization of the information. When a knowledge model is used to annotate data, the occurrences of the defined concepts and relations are added to the structure as instances of these abstract types.

**Ontologies** are probably the most famous type of knowledge models, as they have been used for decades in a number of different domains [Ashburner et al., 2000, Navigli and Ponzetto, 2010, Miller et al., 1990, Kim et al., 2013b]. Ontologies can be defined as «A formal, explicit specification of a shared conceptualisation» [Studer et al., 1998]. In reality, in addition to formal ontologies, the term ontology is used for a number of other representations with varying depths of formality [Guarino and Welty, 2000]. A minimal definition that covers all these scenarios is the following: «An ontology is a specification of a conceptualization» [Gruber, 1993].

In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge [Liu and Tamer Özsu, 2009]. Knowledge is represented and organized in classes (or sets), attributes (or properties), and relationships (or relations among class members). These representations exist on a semantic level and they are intuitive for the human mind. Ontologies are written in formal languages such as RDF-S or OWL with expressiveness close to first-order and modal logics. These languages allow abstraction away from specific structures and tools and integration of heterogeneous data sources.

Even though their purpose is not limited to information extraction applications, knowledge models and ontologies, in particular, are an essential part of IE as they are the foundation on which an IE task is [Nédellec et al., 2009]. They define the domain and scope of the task, as well as the necessary and sufficient set of entity and relation types. They are the guide and the product of the collaboration with the domain expert and provide a reference for the prediction model as well as any later transformation of the extracted data. In the definition of IE tasks, the description of the corresponding knowledge model is always necessary. It is sometimes given explicitly, like in the example of the GENIES system, for which the model was also published [Rzhetsky et al., 2000], [Friedman et al., 2001]. In other cases, it is implicitly described by the definition of the IE task, as it is the case for the historical MUC challenges.

The term **domain knowledge** can be used to refer to the set of information necessary to describe a domain. Domain knowledge is represented as knowledge models (such as ontologies), databases or other structured, formal, machine-readable representations. In the case of expert-produced knowledge bases, the criterion of inclusion or exclusion of concepts and relations in these structures is their contribution to and necessity for the domain description according to the experts. An example of expert-produced models are those created for the BioNLP Bacteria Biotopes and Gene Regulation Network (GRN) challenges [Bossy et al., 2012, Bossy et al., 2015]. In the case of automatic acquisition of domain knowledge, the scope is defined by the choice of input data. Automatic acquisition of domain knowledge is generally based on text [Yangarber et al., 2000, Maedche and Staab, 2001, Buitelaar et al., 2004, Buitelaar and Cimiano, 2008].

Information extraction is often described as an extraction of **structured information** from unstructured text [McCallum, 2005]. The structure of this information is defined by the knowledge model, and the information extracted can be used to populate or enrich the initial model. The extracted information coupled with the context of the model is considered knowledge, but by itself it only satisfies the definition of information given above.

Another related term is that of **knowledge inference**, which is the process of *reasoning*, or using logic rules and constraints in order to obtain new knowledge from already established facts [Tari, 2013]. Whereas knowledge models can be used for inference [Gardner et al., 2013], knowledge inference is a different approach to information extraction, as it is based on logic in order to extend and enrich the knowledge model. Information extraction, on the other hand, uses text and other resources and it discovers new knowledge by aligning them to the given knowledge model.

A knowledge model, finally, is distinct from a text **annotation schema**, the set of concepts and relationships which are used to annotate the textual data, manually or automatically. An annotation schema is generally a selected subset of the knowledge model, related to a specific application and implementation. The choices concerning which concepts and relationships are to be part of an annotation schema are a result of the goal of the application, and not an ambition to fully represent or describe a domain. These choices can be influenced by a number of facts. The specificities of expression is one factor: literary texts might, for example, employ pronouns more often than wiki pages, thus making the annotation of anaphora more important. The annotation process equally plays an important role: for many automated tools as well as manual annotators it is often more straightforward to use relationships having only two arguments instead of more complex constructs.

The annotation schema may also be different from the structured **extracted information**, in the case of different tool constraints and conventions. One such example is the case of the BioNLP Challenge Genia tasks [Kim et al., 2012, Kim et al., 2013a], where the manual annotations are binary relations, but the expected extracted information is in the form of complex events often involving more than two arguments. Another such example is the case of the BioNLP GRN task [Bossy et al., 2013a] where the targeted information is in the form of a regulatory network, whereas the annotations are, again, binary relations.

### 1.4.2 Knowledge Extraction for Biology and *A.Thaliana*

#### Importance

#### From the biologist's point of view

With the emergence of systems biology as a major field of biological sciences, databases and models started playing an important role in biology [Ideker et al., 2003]. Knowledge extraction and literature mining have practical advantages with regards to automatic database and model construction and update, as well as the ease of access to knowledge [Craven and Kumlien, 1999]. From the biologist's point of view, these methods are becoming increasingly useful for both hypothesis generation and biological discovery [Jensen et al., 2006, Craven and Kumlien, 1999], thanks to the growing volumes of text in general and in open-access libraries and journals.

Even though a great number of databases exist, data from biological research are often not submitted in all their detail or even at all to these databases. However, they are always reported in scientific publications, thus making scientific literature a more complete and up-to-date resource of biological knowledge [Gieger et al., 2003, Rebholz-Schuhmann et al., 2005]. Moreover, because of the nature of scientific articles, observations and experiments have the advantage of being accompanied by context information and an interpretation.

### From the computer scientist's point of view

Biomedicine has gathered a lot of attention as a domain of application for knowledge extraction. From the point of view of the knowledge extraction researcher, it is a highly attractive domain thanks to the availability of structured resources such as document collections and ontologies, as well as the eagerness of the biomedical community to explore new approaches. Additionally, applications in these domains have the benefit of a clear requirements and a large amount of evaluation material.

While nowadays large volumes of text are available through the web and other databases, texts in the biomedical domain are particularly adapted for IE tasks as they are usually in the form of reports or published articles which are dense in factual information. Texts of general interest or business interest, for example, can often contain commentary or personal opinions, which require additional processing, such as sentiment analysis. Moreover, large literature collections and open-access publications are available, namely PubMed/MEDLINE, BMC and PLOS.

Additionally, in regards to domain knowledge, information extraction on the biomedical domain (sometimes called BioNLP [Cohen, 2010]) has the advantage of well documented structures and the existence of a great number of databases, like the examples of Gene Ontology [Ashburner et al., 2000] and UMLS [Bodenreider, 2004], or the MAtDB [Schoof et al., 2002] [Schoof et al., 2004] in the cases of Arabidopsis. As scientific fields, they are widely standardized and documented, and as domains tied to the physical world, knowledge is easier to express clearly and organize. While knowledge bases can have different perspectives or ambitions, existing resources can often be the basis of new approaches.

The two most essential conditions for a successful knowledge extraction application are the ease of definition of the model and the availability of data, both of which facts are true in the case of BioNLP. Whereas other domains might satisfy one or both of these conditions, the communities of biology and medicine are by large the most active in their involvement in knowledge and information extraction. Currently, most of the international challenges and workshops in knowledge extraction from text are focused on the biomedical domain [Arighi et al., 2013, Ohta et al., 2013b, Tsatsaronis et al., 2015, Uzuner et al., 2011].

While the biomedical domain as a whole shares a number of common features, it includes very distinct subdomains. Firstly, one major factor of differentiation is the context in which they exist. This is the main reason medical applications are often a special case; the stakes are often more critical and at the same time privacy and access to resources is strictly controlled.

But even within the group of biological sciences, the nature of the information might vary greatly. For example, in the case of molecular biology, sources of information will include high volumes of data produced by laboratory instruments such as sequencers, as well as databases and experiment documentation. In the case of ecology, on the other hand, data will mostly come from databases and publications detailing the discovery of organisms in different environments.

#### A short historical survey of projects and applications

The potential of text mining as an alternative method of accessing knowledge was first explored in the contexts of database curation and scientific information retrieval [Craven and Kumlien, 1999, Eilbeck et al., 1999, Pulavarthi et al., 2000, Tamames et al., 1998, Jenssen et al., 2001, Müller et al., 2004, Hoffmann and Valencia, 2004]. In systems biology [Ananiadou et al., 2006], text has helped parameter learning for models [Hakenberg et al., 2004], it has often been used to make connections between seemingly dissociated arguments [Weeber et al., 2003, Swanson, 1988, Smalheiser and Swanson, 1994, Srinivasan and Libbus, 2004], and in order to add context and interpretation to experimental microarray data [Krallinger et al., 2005, Oliveros et al., 2000, Blaschke et al., 2001, Shatkay et al., 2000, Raychaudhuri and Altman, 2003, Imoto et al., 2011, Faro et al., 2012].

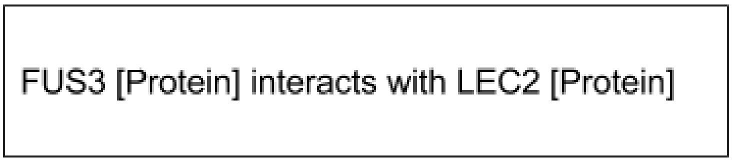
Most of the early BioNLP projects focused on simple interactions between genes and proteins [Blaschke et al., 1999, Nédellec, 2005a, Yeh et al., 2002, Yeh et al., 2003, Hersh and William, 2004]. More recently, the community has been exploring more ambitious goals with more complex extraction tasks, such as the extraction of more intricate biological events [Kim et al., 2012, Kim et al., 2009a, Kim et al., 2003, Kim et al., 2011, Kim et al., 2004], the extraction and reconstruction of networks [Bossy et al., 2013a, Li et al., 2013, Ramani et al., 2005] and pathway curation [Ohta et al., 2013a] tasks.

Historically, the first information extraction projects concerned literature on human, mouse and fly biology [Hirschman et al., 2005, Hersh and William, 2004, Hersh et al., 2006, Hersh et al., 2008, Kim et al., 2003, Ohta et al., 2013a, Ramani et al., 2005]. The LLL challenge [Nédellec, 2005a] was the first to introduce bacterial biology, followed by the BioNLP Bacteria Biotope task [Bossy et al., 2013b, Bossy et al., 2011a]. Plant biology has so far been relatively underrepresented as a topic for the BioNLP community. *Arabidopsis thaliana* has recently seen some initiatives in the field of Information Extraction, such as the KnownLeaf literature curation system [Van Landeghem et al., 2013, Szakonyi et al., 2015]. It is worth noting that there have been other text mining applications on *A. thaliana* in the past, but they were mostly focused on information retrieval [Krallinger et al., 2009, Van Auken et al., 2012].

## 1.5 Knowledge Expressed in Text: the Corpus

### 1.5.1 The building blocks: entities and semantic relations

We use natural language to express and communicate our knowledge. Scientific articles are the standard way to share scientific discoveries, methods and results, to summarize and analyze facts,



FUS3 [Protein] interacts with LEC2 [Protein]

Figure 1.6: Example of a relation in PPI, along with the entities which serve as its arguments.

commentary and conclusions. Along with longer works such as books, they constitute the best way to access scientific knowledge. This knowledge expressed in scientific text can be extracted by a set of automated methods defining the domain of knowledge extraction. The first step is Information Extraction (IE).

Information extraction is the process of identifying within text instances of specified classes of entities and of predications involving these entities [Grishman, 1997]. Its purpose is making the information in the text more accessible for further processing [Grishman, 2012]. This further processing step is the one that provides context and structure, transforming information into knowledge.

The two basic concepts of knowledge and information extraction from text are entities and relations. By **entities** we refer to the parts of text describing discrete entities such as people, organizations, living organisms or their parts, or even objects. Even though texts include a large number of entities, a specific information extraction task targets a predefined set of entities of interest, relevant to the application. For example, in the case of protein-protein interaction (PPI) extraction, the targeted entities are proteins (Fig. 1.6). It is the knowledge model for each task that defines the set of entities to extract. These entities are designated in the text by proper names, nominal phrases and pronouns, *etc.*

A **relation** represents some semantic relationship between entities. In the context of information extraction, we use the term **relations** to specifically designate predications about a pair of entities, and **events** for relations involving more than two entities or relations [Grishman, 2012]. As it is the case with entities, the set of relations and events targeted are defined by the accompanying model and their nature and meaning depend on that specific knowledge domain. Models for extraction tasks vary greatly in their complexity, ranging from the PPI single-relation model (below) to models involving multi-argument events such as the Genia model [Kim et al., 2003]. Relations and events can be affirmations or negations, they can have conditions or even express speculation.

### 1.5.2 Corpus

A textual **corpus** is a collection of texts put together for a specific purpose. A corpus can include documents in different languages or of different natures and formats, *e.g.* newspaper article, wiki page, book chapter. In the context of information extraction, corpora are almost exclusively in one language, and –more often than not– of one specific type.

### Corpus Design *&* Annotation

An annotated corpus is defined by its schema and its annotated documents. The decisive factor in corpus design is the purpose a corpus is intended to be used for [Lüdeling, 2008]. Three approaches of corpus design can be defined [Fort, 2011]: a) top down, where the focus is on the knowledge model, b) bottom-up where the text is central in the design and c) a mixed approach of cyclic annotation. This third approach uses annotation iterations where the theory (model) is tested on real data and redefined as necessary.

Corpus annotation is defined by Leech in [Garside et al., 1997] as the practice of adding interpretative, linguistic information to an electronic corpus of [spoken and/or] written language data. The end-product of this process is often referred to as annotation(s), too. The purpose of annotation is to enhance raw data with linguistic annotations [Bird and Liberman, 2001] relevant to a specific task, while at the same time ensuring reusability [Pogodalla, 2009].

Leech [Leech, 1993] proposed seven maxims which summarize the written and unwritten rules of good corpus annotation:

1. It should always be possible to come back to initial data.

2. Annotations should be extractable from the text.

3. The annotation procedure should be documented.

4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)

5. Annotation is an act of interpretation, and as such it cannot be infallible.

6. Annotation schemas should be as independent as possible on formalisms.

7. No annotation schema should consider itself a standard, even though it can possibly become one.

The quality and sophistication of the annotation is integral to the (re)usability of a corpus, or as Wallis [Wallis, 2007] put it: “you only get out [of the corpus] what you put in”. Still, annotation campaigns involve lengthy and costly manual work, and are demanding in technical aspects. The creators of the BioInfer corpus report 15 man-months of annotation efforts for a corpus of 1100 sentences [Pyysalo et al., 2007].

In most cases, the annotators are not experts of the annotation as a process, but of the application domain of the corpus. Good training is considered the best way to improve the speed and the quality of annotations [Marcus et al., 1993, Chamberlain et al., 2008, Dandapat et al., 2009]. Adapted documentation –generally called “Annotation Guidelines” or “Annotation Guide”– is equally essential, providing a clear definition of the application, the entities and relations, giving meaningful examples and covering any possible ambiguities [Fort, 2011]. The quality of the guidelines has a direct effect on the performance of any machine learning method for Information Extraction [Nédellec et al., 2006].

While the popularity of a corpus is closely related to its domain of application, several *measures* exist in order to quantify the quality of the annotations, like kappa statistics and correlation coefficients. They measure inter- and intra-annotator agreement. Since annotation is optimally done by more than one annotators, inter-annotator agreement takes into account the differences of annotations between them. In the spirit of an iterative corpus design, [Bonneau-Maynard et al., 2005] suggest computing inter-annotator agreement early in the campaign and subsequently making necessary updates in the guidelines. As campaigns often last several months, it is crucial that the coherence of each individual annotator's annotations are often checked [Gut and Bayerl, 2004]. This intra-annotator agreement is ensured by a good definition of the task and a good guide.

During an annotation campaign, these measures can be guides for the revision of the schema and the guidelines. In practice, annotation campaigns and corpus design are often done in an iterative manner, with approaches going as far as Agile annotation [Alex et al., 2010, Voormann and Gut, 2008].

**Annotation Editors** are the software tools used in the annotation process. They display the text and provide a usual interface to add entity and relation annotations. Traditionally, annotation editors are standalone programs (GATE [Cunningham, 2002], Glozz [Widlöcher and Mathet, 2012], Knowtator [Ogren, 2006], etc), but recently a number of web-based annotation editors have been developed, improving interoperability and collaborative editing such as Brat [Stenetorp et al., 2012], AlvisAE [Papazian et al., 2012], and TextAE [Kim et al., 2015].

Finally, alternative approaches such as using crowdsourcing for annotations have been used and evaluated [Fort et al., 2011, Saunders et al., 2013]. Crowdsourcing has been found to provide an effective alternative to traditional annotation, but it can have limitations, generally related to the expertise necessary to annotate, and the quality of the process insured by each platform.

### General Purpose Corpora

As the choice of corpus reflects the type of knowledge or language one wishes to model, the majority of corpora used for most natural language processing (NLP) and linguistics applications are based on text originating in “general interest” sources, like news articles. Shared corpora and international challenges play a major role, as they increase focus and facilitate rapid advances in the field [Pyysalo et al., 2007].

The Message Understanding Conference (MUC) [Grishman and Sundheim, 1996] competitions marked the beginning of IE as a domain and contained text coming initially from military reports and later from newswire and journal articles [LDC, 1993, LDC, 2001]. The task invited participants to propose approaches to extract information on fleet information, terrorist attacks, airplane crashes and business data from these messages.

The Automatic Content Extraction (ACE) [Doddington et al., 2004] program also followed a similar approach, basing the corpus on text coming from the news. ACE entities included person, organization and location, among others. Relations covered physical relations such as location and affiliations (business, organizations, ethnicity etc).

The CoNLL challenge covers a number of NLP topics and has included IE-related tasks in the past. It has often used news articles as a source, namely sections from the Penn Treebank Wall Street Journal (WSJ) corpus [Marcus et al., 1993] and the Reuters corpus [Lewis et al., 2004], as well as parts of the Linguistic Data Consortium ECI corpus [Linguistic Data Consortium, 1994], and the Ontonotes Corpus [LDC, 2011a] from the LDC. The Ontonotes Corpus has also been used in other works, such as SemEval 2010 [LDC, 2011b] task on coreference resolution.

Other corpora of general interest adapted to information extraction include the New York Annotated Corpus [LDC, 2008a], the Wikipedia-based Google Relation Extraction Corpus [Google, 2013] and the WikiLinks cross-document coreference corpus [Singh et al., 2012]. Finally, the LDC Catalog [LDC, 1992] includes a large collection of available corpora, categorized by source type and application.

### Corpora Specialized in Biology

#### Corpus design & annotation for biomedical corpora

News text and wikipedia articles are used as sources for most general interest extraction tasks. However, they are not appropriate for biology, as they do not correspond to the knowledge of the domain. Wikipedia articles on biology topics could be possible candidates, but in practice scientific text is considered richer in information and detail.

Publicly available corpora are often presented and distributed in the context of challenges. Depending on the task, they can come with additional pre-annotation, for example syntax, coreference or entities. This pre-annotation is shared in order to evaluate each sub-task of IE separately. The language used in biomedical scientific text is different in nature, both in vocabulary and style, thus making standard general corpora of limited utility for BioNLP [Pyysalo et al., 2007]. For this reason, standard linguistic tools used in corpus preprocessing for IE often need to be adapted or re-trained on domain specific corpora [Kim et al., 2008].

Both model design and annotation need to be done in collaboration with experts of the target knowledge domain. Multidisciplinary collaborations call for a well-thought strategy as each partner has specific time constraints, experimental limitations, and sociological basis [Krallinger et al., 2009].

Creating a representative collection is a critical step for the success of an IE task. Consequently, the selection of documents is a very important part in corpus design. Krallinger et al. [Krallinger et al., 2009] identify four methods for automated selection: 1) based on biological database citations and links [Blaschke et al., 1999], 2) based on a select choice of journals which are known to have good quality standards, 3) a term-mapping approach using keywords and terms and 4) an entity-mapping approach using entities coming from an appropriate knowledge base. An alternative approach relying on knowledge and experience is manual selection by the domain experts. All approaches bear the risk of biasing the corpus, and one can argue that manual selection is particularly prone to this shortcoming, but it has the advantage of ensuring a good quality and coverage of the knowledge domain. A mixed approach combining automated collections and manual filtering can be used to try and have the best of both worlds.

Despite the fact that corpus design and annotation are tedious and time consuming tasks, well-organized campaigns can produce reusable and reliable corpora, such as the example of the popular Genia Corpus [Kim et al., 2008]. For this corpus of 400,000 word annotations from scientific abstracts the work of a group of 7 experts for 1.5 year was necessary. This team included 5 part-time annotators, 1 senior coordinator and 1 junior coordinator [Kim et al., 2012, Kim et al., 2009a, Kim et al., 2003, Kim et al., 2011, Kim et al., 2004].

Even though the focus of this section is corpora specifically adapted to IE, it is worthy to note that there exist a number of well known biomedical corpora destined to be used for information retrieval, such as the TREC Genomics corpora

[Hersh et al., 2006, Hersh et al., 2008, Roberts et al., 2008, Hersh and William, 2004] and the KDD Challenge Task 1 (2002) [Yeh et al., 2003, Yeh et al., 2002] corpus.

Initially, most biomedical corpora consisted mostly of isolated sentences originating in publication abstracts [Cohen et al., 2005]. More recent corpora have tried to include longer texts. For example, the Bacteria Biotope corpus [Bossy et al., 2011b] uses complete web pages and wiki articles, and the CRAFT corpus [Bada et al., 2012] consists of full-length scientific articles. The type of extraction is another variable in corpora, with entity extraction being the main focus of earlier works, while more complicated tasks are becoming increasingly common.

### Biological Corpora Themes

Regarding the choice of topics for corpora on biology, **regulation** has been the central topic of most corpora and IE works. The regulatory relation around which a great number of corpora have been built is the interaction. Protein-protein interaction (PPI) and gene-protein interactions are the most common examples.

The first PPI corpus appeared in 1999 [Blaschke et al., 1999] and consisted of selected MEDLINE abstracts. One of the first approaches in automated interaction extraction, it contained no manual annotations. The PennBioIE corpus [Kulick et al., 2004, LDC, 2008b] consists of 1.414 MEDLINE abstracts on cancer molecular genetics, citing 18.148 entity annotations.

The Biocreative I [Hirschman et al., 2005] challenge included two tasks : task 1 focused on extraction of gene mentions from single sentences in MEDLINE abstracts and associating database identifiers. Task 2 focused on identifying text passages in full text articles that provide evidence for GO annotations about a particular protein. The BioInfer Corpus [Pyysalo et al., 2007] focused on the development of IE systems for extracting relationships between genes, proteins, and RNAs and included 1100 sentences from abstracts. A more recent corpus that focuses entirely on entities is the CRAFT corpus [Bada et al., 2012].

As mentioned earlier, with most biology corpora being almost exclusively about human and animal biology, the LLL corpus [Nédellec, 2005a] was different in that aspect as it was based on texts on bacteria, while also focusing on interactions. The GRN [Bossy et al., 2013a] corpus for the BioNLP challenge continues in the same direction but aims to extract regulatory networks instead of isolated relations. The Pathway Curation (PC) [Ohta et al., 2013a] task of the same challenge is also aiming for more sophisticated regulatory knowledge extraction, targeting pathways.

The BioNLP Bacteria Biotope (BB) corpus [Bossy et al., 2013b, Bossy et al., 2011a] shares with LLL and GRN the bacteria theme, but approaches the subject from a different perspective. Instead of regulation knowledge, the BB corpus consists of texts describing bacterial habitats.

Another theme that has gathered a lot of interest in biomedical IE is the association of genes and phenotypes, and diseases in particular. Examples of corpora in this domain include the gene-disease corpus of Chun et al. [Chun et al., 2006] who created a corpus of 1000 manually annotated sentences, selected from MEDLINE abstracts by using co-occurrence and MeSH terms. The EUD-AD corpus [van Mulligen et al., 2012] contains 300 MEDLINE abstracts fully annotated by three annotators with relations between drug, disorder, and targets.

While corpora on plants and *Arabidopsis thaliana* in particular have been published before, the applications have mostly been limited to information retrieval [Van Auken et al., 2012] and the building of knowledge bases [Krallinger et al., 2009].

The ATCR (*Arabidopsis Thaliana* Circadian Rhythms) Corpus [Rinaldi et al., 2007], although consisting of automatically annotated text, is one of the first examples of knowledge mining corpora for *A. thaliana*. The potential of knowledge extraction from literature and the combination of this knowledge with existing sources have been gathering interest in the last years [Van Landeghem et al., 2013].

The recent KnownLeaf project [Szakonyi et al., 2015] on *Arabidopsis* focuses on the regulatory mechanisms in leaf growth and development and integrates 283 research articles annotated by multiple annotators. The resulting database of 9947 relations was used to produce a graphically represented network of extracted knowledge.

## 1.6 Information Extraction

### 1.6.1 Defining IE

#### What IE is

Historically, Information Extraction emerged as a domain of research as a result of the the DARPA MUC program (Message Understanding Conference) [Chinchor et al., 1993, Lewis, 1991]. The final MUC-7 conference defined it by distinguishing five separate tasks: 1) Named Entity Recognition (NE), which detects and classifies entities such as names, places, etc., 2) Coreference resolution (CO), which finds identity relations between entities, 3) Template Element Construction (TE), which fills templates with descriptive information about entities, 4) Template Relation Construction (TR), which finds relations between entities, and 5) Scenario Template Production (ST), which combines TE and TR results and fits them into specified event scenarios [Cunningham, 2006]. The ACE (Automatic Content Extraction) workshops that followed MUC also grew to include tasks of entity, relation and event extraction.

Since the MUC conferences the view on IE has shifted from being template-based to being model-based. Consequently, the terms used to describe the subtasks no longer reference templates. The prevailing terminology today, used in this thesis, describes the following parts of IE:

- Named Entity Recognition (NER)
- Coreference Resolution, which additionally often is combined with grammatical anaphora resolution.
- Relation Extraction (RE), for relations between pairs of entities
- Event Extraction, for more complex events implicating more than two entities and, possibly, relations or other events.

Grishman [Grishman, 2003] provides a definition of IE as the process of identifying and classifying instances of some sort in text, based on some semantic criterion. Going even further in the direction of knowledge extraction, Nédellec et al. [Nédellec et al., 2009] define IE as a process that aims at mapping text to ontology. For this definition, in an IE system the relevant pieces of input text are selected and interpreted using semantic analysis, according to an ontological knowledge structure.

### What IE is not

Even though related, Information Extraction and **Information Retrieval** (IR) are two different tasks. An IR system finds and retrieves relevant texts and presents them to the user, where a traditional IE system analyzes texts and presents only specific, interesting, information from them [Cunningham, 2006].

**Question Answering** (Q/A) is closer to IE than IR in the sense that both tasks try to find in text the pieces of information that are relevant to a given query. However, Q/A systems answer a wide range of unpredictable user questions, whereas an IE system looks for richer but predefined type of information [Nédellec et al., 2009]. It is worth noting that IE has been considered as an approach for Q/A [Srihari and Li, 1999].

Keeping in mind the different ambitions of these applications, IE systems that share some similarities with IR emerge as IE moves away from a template-based view. A comparison to illustrate this point can be made between the BioCreative IR tasks and the BioNLP GRN network. In the former, the IR systems seek to detect whether a text matches, without identifying the specific part of the text, or the structured information as it is in the text. In the latter, the existence of specific, structured information is used to infer a network, and it is over this network that the systems are evaluated and not the originating parts of text.

Lastly, traditional IE is different than **Open IE.** Where traditional IE deals with homogeneous corpora and precise, predefined models, OpenIE is relation-independent and tailored to massive and heterogeneous corpora such as the Web. An Open IE system extracts a diverse set of relational tuples from text without any relation-specific input[Banko et al., 2008]. Another difference of IE from the Web is that web pages in addition to text can include structured information in the form of HTML tags, document structure, *etc.* [Kayed et al., 2006].

### 1.6.2 Evaluating IE systems

International challenges (or shared tasks) have since the beginning played a very important role for IE. They have provided structure and motivation for the development of new approaches and a means of evaluation and comparison of IE systems.

A non-exhaustive list of IE-related challenges, with emphasis on recent tasks, is the following:

*Biomedical:*

- TREC Genomics: 2003 - 2007
- BioCreative: 2004 - 2015
- i2b2 Challenges in Natural Language Processing for Clinical Data: 2008 - 2014
- BioNLP ST : 2009, 2011, 2013, 2016
- Clef eHealth Information extraction from Clinical Text: 2013, 2014, 2015, 2016

*Other*:

- MUC (Message Understanding Conference) : 1987 - 1997
- ACE (Automatic Content Extraction): 2002 - 2008

- LD4IE (Linked Data for Information Extraction): 2014, 2015
- COLIEE (Competition on Legal Information Extraction/Entailment): 2014, 2015

While shared tasks like the above provide a framework for evaluation of IE systems, they generally suffer from two main setbacks. The first one is overfitting and is inherent with small standardized tests and small datasets. While challenge organizers take measures to avoid overfitting such as providing separate development and test datasets, competing systems inevitably optimize their performance on these particular tasks. Shared tasks are organized to be easy to analyze and study, with subtasks being frequently proposed separately. Consequently, good performance on a challenge is not guaranteed to translate into equally good performance in larger real-life IE tasks.

The second setback has to do with the choice of evaluation metrics. Standard metrics such as precision, recall and F-measure are a default choice, thanks to their intuitivity and popularity in the Machine Learning field. While these measures correspond well to the evaluation of ML techniques, they are not always adapted to IE. An example that illustrates this is the following: in relation extraction, a substitution error is a prediction that has the correct arguments but an incorrect label. A substitution error is counted twice by evaluated systems, as a false positive and a false negative. In the calculation of the F-measure, a substitution is penalized twice, overestimating the deviation from the reference. Fortunately, this setback is not inevitable and recent tasks have introduced new evaluation measures which are more adapted to IE, such as the Slot Error Rate (SER) [Galibert et al., 2010] used in the BioNLP GRN task [Bossy et al., 2015].

### 1.6.3 IE Systems

#### Overview: Two axes of evolution

From the early template-filling approaches to the sophisticated complex network extraction techniques of tomorrow, IE systems have evolved along two main axes: the level of linguistic information and pre-processing they use, and their prediction algorithms.

Following the early **pattern-matching** based systems such as Autoslog [Riloff, 1993], the first systems for biomedical IE used similar approaches –for example [Ng and Wong, 1999, McDonald et al., 2004]– to extract pathway information. While rules and regular expressions can be very precise predictors, they fall short in covering the breadth of expression in natural language text.

**Pattern learning** tries to overcome the problem of hand crafting patterns by introducing a learning procedure [Huffman, 1995, Agichtein and Gravano, 2000, Brin et al., 1998]. Going even further, by introducing linguistic pre-processing and syntactic analysis [Park et al., 2001] and Yakushiji et al. [Yakushiji et al., 2001] propose pattern approaches based on a higher abstraction layer than lexical representations.

At the cost of reducing accuracy, **co-occurrence** approaches [Fukuda et al., 1998, Stapley and Benoit, 2000] for relation extraction can address the issue of variability in expression. The reason behind this is that simple co-occurrence of two entities does not guarantee the relevance and type of semantic relation between them. While a combination of co-occurrence with specific words (often called trigger words) have been used to filter co-occurrences, invariably, linguistic information is obligatory in order to normalize and generalize in a meaningful way. These trigger words are words (or phrases) of which the occurrence in a sentence is an indicator that this sentence includes a semantic relation. These words are generally specific to one relation type. Producing a representative and exhaustive set of trigger words is a demanding task, often penalizing the recall performance of the prediction. Some methods use machine learning to predict them. But even with the use of such words, syntactic relations are a much greater predictor of semantic relations than proximity in the context sentence.

However, if no **coreference resolution** is done, IE systems are limited to extracting relations that occur exclusively within a sentence. Two of the first systems which included anaphora resolution were those of Pustejovsky et al. [Pustejovsky et al., 2002] and Stapley et al. [Stapley et al., 2002].

**Machine learning** provides an alternative to hand-crafted patterns and has been employed in IE for many years. Many different approaches have been used for all of the IE subtasks, including Hidden Markov Models (HMM) [Freitag and McCallum, 1999, Skounakis et al., 2003], Conditional Random Fields (CRFs) [McCallum, 2002, McDonald and Pereira, 2005] for NER, maximum-entropy models [Kambhatla, 2004] and kernel-based methods [Bunescu and Mooney, 2005] for RE. With such a rich ecosystem of methods used, overview articles presenting and comparing the state-of-the-art are frequent in IE [Jensen et al., 2006, Hirschman et al., 2002b, Bretonnel Cohen and Hunter, 2004, Yandell and Majoros, 2002, Ananiadou et al., 2006, Zweigenbaum et al., 2007, Tikk et al., 2010].

### Named Entity Recognition

A Named Entity (NE) is an entity or phrase of interest for the task of IE, as defined by the given schema of a specific IE task. Named Entity Recognition is the task of identifying in the text and extracting the Named Entities defined by the model. This includes detecting the existence of a NE and finding correct textual boundaries, as well as its classification as the correct NE type. NER has been the most intensively studied and applied IE task [Grishman, 2003].

For NER, Grishman [Grishman, 2012] cites the following families of methods used today: hand-coded methods, supervised methods (such as CRFs or sequence models), semi-supervised methods and active learning. Most general-purpose tools, such as the Stanford NLP NER [Finkel et al., 2005a] identify three basic classes of NEs: persons, organizations and locations.

NER in the biomedical domain includes some simpler cases. For example, for specific extraction tasks such as the names of organisms dictionaries are often used [Ono et al., 2001][Hirschman et al., 2002a]. Rule-based approaches have equally proven useful for such tasks [Fukuda et al., 1998, Narayanaswamy et al., 2003].

However, biomedical NEs can also be particularly complex, as they are created and used by many different communities [Ananiadou et al., 2004]. A great number of variants and synonyms exist and they can also often be ambiguous. Their ambiguity can be due to terms being employed interchangeably with different meanings as is the case with gene and protein names, or it can be due to biological terms having names which are shared with common English words (an, and, “leaf rust”, *etc.*). Additionally biological named entities are often ambiguous in their boundaries [Zweigenbaum et al., 2007]. Machine-learning has been used to adapt traditional approaches [Tsuruoka and Tsujii, 2004, Yeganova et al., 2004] or to perform NER from scratch [Kazama et al., 2002, Yamamoto et al., 2003, Zhou et al., 2004, Morgan et al., 2004, Collier and Takeuchi, 2004].

Adapting techniques to the biomedical domain is necessary, as studies have showed that although any domain can be reasonably supported, porting a system to a new domain or textual genre remains a major challenge [Nadeau and Sekine, 2007].

In international challenges often IE tasks are split into subtasks separating NER and RE. RE in this case is performed on given NEs. However, in real world applications both tasks are necessary and can be performed in parallel or sequentially.

### Coreference Resolution

Coreference in the context of IE describes the case where two words, phrases or, generally, parts of text refer to the same NE. We talk about coreference resolution when we are looking at pronouns or nouns (anaphors) that refer to a NE (referent or antecedent). NEs can also be anaphors, in the sense that different occurrences detected in text can refer to a single entity, in which case these entities need to be linked.

For the detection of candidate anaphors there are different approaches, based on their nature. For grammatical phenomena such as anaphora, cataphora, etc, for example an approach would look at pronouns in particular. For nouns, the simplest approaches include dictionaries of words that are specific to each corpus.

Research in coreference resolution focuses in ML-based approaches [Ng and Cardie, 2002, Soon et al., 2001, Su et al., 2008]. Nevertheless, rule-based or mixed approaches are widely used. Additionally, external resources such as dictionaries or ontologies can also be used to enrich rules, especially in the case of the biomedical domain (*eg.* [Lin et al., 2004]). The Genia task of BioNLP 2011 proposed a protein coreference challenge [Kim et al., 2012], with two out of the six participants using rules, two machine learning and two combining rules and ML.

Coreference resolution and entity linking [Rao et al., 2013] are essential for Information Extraction. Grishman defines entity extraction as both identifying and classifying entity mentions, as well as linking entity mentions which refer to the same entity [Grishman, 2003, Grishman, 2012], illustrating the importance and necessity of CR. Gabbard et al. [Gabbard et al., 2011] further demonstrate the importance of CR on relation extraction, by showing significant improvements in RE F-scores, even when the CR has high error rates.

Still, CR is often omitted in IE systems. A recent example is the participation in the BioNLP ST BB task, where RE results were found to be heavily impacted by the frequence of anaphora [Bossy et al., 2013b], as none of the initial participants in the RE task performed anaphora resolution. In the aforementioned Genia task participants were invited to submit solutions to both event extraction and coreference resolution. It is noteworthy that of the 12 teams participating in the extraction subtask, only two also submitted solutions for coreference-resolution.

### Relation Extraction

Relation Extraction is the detection and classification of relations between NEs. By relation extraction we refer to the extraction of binary relations, or relations having two arguments, both of which are either NEs or anaphors. As opposed to relations, the term *event* is used in order to refer to more complex relationships between more than two arguments, or those which can take another relation as an argument.

The level of **linguistic information** used can vary greatly and have a direct impact on the performance of a RE system, both in regards to accuracy and recall of the extraction, but also in its computational performance. Most systems aim at higher extraction performance, due mainly to the fact that large-scale real-time applications are rare in the domain of traditional IE[1].

The levels of linguistic information used for RE are: surface (text), syntax & grammar, semantics. Standard pre-processing includes the following:

1. **Tokenization/segmentation** is separating text into words and phrases, and can influence the results in the examples of words with hyphens, numerals etc. For biological text, special well-thought rules are required.
2. **Lemmatization/normalization**, is the process of attaching a canonical form to each word, this can mean the infinitive in the case of verbs, or the singular for nouns, but in biological text it is often coupled with finding the normalized form of specific entities.
3. **Syntactic parsing** produces a syntactic tree for each sentence and can be generally either dependency-based or constituent-based.

Another aspect for examining relation extraction methods is whether they use Machine Learning, and when they do whether they prefer feature or kernel based approaches. While earlier systems often relied on hand-written patterns and rules or co-occurrence, the vast majority of modern RE systems use some kind of Machine Learning for the extraction of relations. Whether it is pattern learning, feature or kernel based models, supervised learning is the standard for RE. However, as annotated data are scarce, a number of semi-supervised approaches have been proposed throughout the years.

### Kernels versus Features : Representation versus Algorithm

A common representation for many machine learning algorithms are feature maps, or feature vector representations. These vectors are n-dimensional vectors of numerical values, where each dimension (or feature) represents a measurable property or observation. The process of creating such features, selecting and combining them in order to improve a ML system is called feature engineering.

Feature-based approaches are very popular for RE, with some recent examples being: [Özgür and Radev, 2009, Fayruzov et al., 2009, Reza et al., 2011, Liu et al., 2012, Kambhatla, 2004]. Feature engineering is most often done manually. Kambhatla [Kambhatla, 2004], for example manually constructs a limited set of features combining various syntactic sources for use with a Maximum Entropy classifier. Crafting such features can be a tedious process, but evaluating and selecting the most useful features is also a difficult task. Fayruzov et al. [Fayruzov et al., 2009] study the linguistic features used for Protein-Protein Interaction extraction and find that only a small subset or the features typically used are actually necessary.

A different family of machine learning methods, called kernel methods, do not require feature engineering as they are based on similarity functions. These functions calculate the pairwise similarity between two instances and thanks to a method called the “kernel trick” they do not

[1] Web and Open IE which have different algorithmic constraints will not be covered by this section.

require a feature vector representation[2]. The kernel trick takes its name from kernel functions. These functions allow operating in an implicit feature space without ever computing the exact coordinates of data in that space but rather by simply computing the inner products between the images of all pairs of data in the feature space.

A shift towards using kernel methods can be observed in recent years; for example all of the approaches and namely the best ranking ones in the Drug-Drug Interaction (DDI) Extraction SemEval 2013 Challenge used kernel methods [Segura Bedmar et al., 2013]. A representative example of this category will be presented below, with regards to the types of linguistic information they use, since they are most often used with syntactic graphs. Additionally, readers are invited to consult Tikk et al which performed a benchmark of kernel methods for PPI extraction in 2010 [Tikk et al., 2010].

### Syntactic Information

The link between syntactic relations and semantic ones is intuitive and I consider the use of the former to predict the latter to be the best approach. This thesis was also explored by Bunescu and Mooney who state that the extraction accuracy increases with the amount of syntactic information used [Bunescu and Mooney, 2005].

However, while deeper representations promise better generalization and semantic relevancy, they inevitably suffer from errors in computing these representations [Zhao and Grishman, 2005]. The direction to take in developing RE systems is the optimization of the use of syntactic information, while taking care to take advantage of shallow information in order to avoid missing information because of errors.

The first learning approaches were those meant to facilitate the production of extraction patterns [Huffman, 1995, Agichtein and Gravano, 2000, Brin et al., 1998], which rapidly evolved to take advantage of syntactic information [Park et al., 2001, Yakushiji et al., 2001, McDonald et al., 2004]. Different approaches avoiding using any syntactic information include treating relation extraction as a sequence labeling task [Culotta et al., 2006].

Approaches using minimal syntactic analysis include the HMM system of Ray and Craven [Ray and Craven, 2001], using just Part-Of-Speech tagging. Going a bit further, shallow parsing (or chunking) has been a popular choice in the kernel-based approaches.

Going again in the same direction, Pustejovsky et al. [Pustejovsky et al., 2002] used shallow parsing and sophisticated anaphora resolution. Zelenko et al. [Zelenko et al., 2003] and Mooney et al. [Mooney et al., 2006] proposed kernel-based approaches on shallow parse trees which gathered a lot of attention, with the latter using them as a sequence, in an approach which reminds of the system of Culotta et al. [Culotta et al., 2006]. Shallow parse tree kernels have seen continued use [Claveau, 2013, Segura-Bedmar et al., 2011] and have been shown to outperform fuller parses in some cases [Giuliano et al., 2006], confirming the hypothesis that parsing errors which occur more frequently in fuller parsers can have a significant impact in the result of RE.

Nevertheless, the IE community invests itself continuously, and for good reason, in the use of deep representations. A number of approaches using parse trees with graph kernels have appeared since Culotta and Sorensen modified Zelenko's shallow parse tree kernel in

[2] In practice and in spite of feature vector representations not being necessary, they are still very often used in combination with kernel methods.

their work on them [Culotta and Sorensen, 2004]. Notable works using graph kernels include [Zhao and Grishman, 2005, Fundel et al., 2007], but parse trees have also been used with convolution kernels [Zhang et al., 2008, Zhang et al., 2006] or even non-kernel methods such as inference [Manine et al., 2009, Manine et al., 2008], or feature transformations [Liu et al., 2007].

Composite kernels making use of both deeper and shallower representations have been used in order to get the best of both worlds: full parses and useful information potentially missed by deep processing [Zhao et al., 2004, Zhao and Grishman, 2005, Zhang et al., 2006].

When using graphs such as parse trees, kernel or feature design which correctly captures semantic information is a difficult task. Bunescu and Mooney [Bunescu and Mooney, 2005] first formulated the hypothesis that the shortest path between candidate relation arguments contains the most relevant information for RE, called *the shortest path hypothesis*. Reducing the complexity of a graph and focusing any method on only the most relevant parts can be very helpful when building RE systems.

Once these (candidate) semantic relations are seen as paths, a number of new interesting approaches can be devised. Airola et al. [Airola et al., 2008] use a graph kernel based on paths from different types of representations. Zhao and Grishman [Zhao and Grishman, 2005] use composite kernels based among other things on dependency paths. Erkan et al. [Erkan et al., 2007] were the first to treat this dependency path as a sequence and apply an edit distance similarity measure on them. Edit distance had been used previously for relation extraction –on Chinese text [Che et al., 2005], but it was solely based on string edit distance.

The edit distance problem was shown to be equivalent to global alignment [Sellers, 1974] and has long been used in string sequence similarity calculations with various applications. Sequence alignment has found applications outside of biology. For example, the Needleman-Wunsch [Needleman and Wunsch, 1970] algorithm has also been adapted for social sciences, where it is known as optimal matching. When aligning two sequences, an element can be aligned with another element, or with a gap. The gap in global alignment is the equivalent of a deletion in edit distance.

Edit distance and alignment methods are great candidates for a dependency path-based kernel approach for RE. Philiipe Veber of INRA's MaIAGE lab developed a first version of such an algorithm [Veber et al., 2011], similar in spirit to Erkan's approach, but with a very important difference: instead of using a true-false comparison between parts of the path, he chose to introduce granularity by imagining a similarity function which gives a non-binary response. Using such a similarity function opens the door to the integration of different types of linguistic information, notably semantics.

### Semantic Information

Different types of information are used in order to extract semantic relations. Using semantic information requires using an external tool or source, for example a dictionary, an ontology, or a lexical database such as WordNet [Miller et al., 1990]. Sources which are created and curated manually contain very accurate information. While WordNet has been used for RE [Claveau, 2013, GuoDong et al., 2005], it suffers from a more general problem with curated resources: the work necessary to build them is tedious and time consuming, resulting in them having a limited scope. Additionally, such sources are often either tailored to a specific domain, or, on the contrary, too general to sufficiently cover the vocabulary of the biological domain.

Domain-specific, curated resources such as ontologies, are rare. In fact, relation extraction is often used to produce ontologies, instead of the opposite [Schutz and Buitelaar, 2005, Buitelaar and Magnini, 2005]. Still, ontologies have been used for RE in the past [Daraselia et al., 2004, Abulaish and Dey, 2007], and will probably be used more frequently in the future as such resources continue to be developed.

On the other extreme, distributional semantics approaches, as statistical methods, can lack the accuracy of human-made collections. However they bring the promise of easy domain adaptation–they can be trained on user-provided text, and richness–large input volume and scope can guarantee a good vocabulary coverage and representativeness. Even though efficient distributional semantics methods such as word embeddings are recent advancements [Mikolov et al., 2013a, Mikolov et al., 2013b, Pennington et al., 2014], there already have been integrations into RE [Gormley et al., 2014, Nguyen et al., 2015] and they will most certainly continue to be explored in the near future.

### Other approaches

Feature-based or kernel-based, most RE systems are classification systems and a vast majority among them use Support Vector Machines (SVMs) as a classifier. Nevertheless, with RE being actively researched for about two decades, there has been a variety of approaches which have been proposed and are not covered by this section. For example some recent examples include: Thomas et al. use Ensemble Learning [Thomas et al., 2011], Barnickel et al. use Neural Networks and Semantic Role Labeling [Barnickel et al., 2009], Khambatla [Kambhatla, 2004] and Mintz et al. [Mintz et al., 2009] use logistic regression classifiers.

Going even further from what was presented, Miwa and Sasaki considered performing NER and RE jointly in their 2014 work [Miwa and Sasaki, 2014], Surdeanu and Tibshirani treat RE as a multi-class, multi-label problem [Surdeanu and Tibshirani, 2012] and Li et al. [Li et al., 2011] as well as Liao et al. [Liao et al., 2010] try to use cross-document information to enhance performance.

Finally, as a last word on Machine Learning for RE, it is worth mentioning that in order to address the issue of the scarcity of manually annotated corpora for classification, semi-supervised or weakly supervised [Riedel et al., 2010] or even unsupervised [Grishman, 2012] approaches have been explored and have recently gathered interest.

### Event Extraction

Hirschman et al. [Hirschman et al., 2002b] identified three subtasks of biomedical IE: NER, RE, and Event extraction (EE). Event extraction differs from relation extraction in two ways.

First, in the way these two tasks are formulated conceptually: relation extraction relies on the notion of relation arguments, as it is defined as the detection and classification of semantic relations between specific arguments. Event extraction, on the other hand, is defined as the detection and classification of biological events, regardless of arguments.

The second way RE and EE differ is closely related to these concept definitions and it concerns the practical way they are defined in extraction systems and, consequently, extracted. Relation extraction is used as a term to describe binary semantic relation extraction, *ie.* the extraction of semantic relations between pairs of arguments. Event extraction, on the other hand, can have a variable number of arguments.

While event extraction is not within the scope of this thesis, the interested reader is invited to start exploring the related bibliography by the following works: [Yakushiji et al., 2001, Kim et al., 2009b, Ananiadou et al., 2010, Kim et al., 2011].

## 1.7 The Alvis ecosystem

**Alvis Suite** is a generic text-mining pipeline based on linguistic and machine learning technologies. It can be easily configured for specific domain applications. Its major components are listed in this section. Figure 1.7 contains an illustration of the suite's architecture.

**AlvisIR** (Alvis Information Retrieval)[3] [4] is an online, generic, semantic search engine. Given a document collection and an ontology, it can create in a few hours a new instance of a semantic search engine. Users can query AlvisIR with the ontology concepts and retrieve all documents that contain the concepts, in the form of specific concepts, or synonyms. AlvisIR also handles relation queries. For example, search on biotopes of microorganisms.

**Alvis NLP/ML**[5] [Nédellec et al., 2009] is a pipeline for the semantic annotation of textual documents. It integrates Natural Language Processing tools for sentence and word segmentation, named-entity recognition, term analysis, semantic typing, and relation extraction (**AlvisRE**, detailed in chapter IV). These tools rely on resources such as terminologies or ontologies for the adaptation to the application domain. New components can be easily integrated into the pipeline.

**AlvisCrawler**[6] is a tool Alvis NLP/ML can use for (semi)-automatic acquisition of the necessary resources. It can automatically download relevant full-text articles from the Web from keyword queries.

**AlvisAE** (*Alvis Annotation Editor*)[7] [Papazian et al., 2012] is an online annotation editor designed to display and edit fine-grained formal semantic annotations of textual documents. It facilitates the collective edition and the visualisation of annotations of entities, relations and groups. It includes a workflow for annotation campaign management. The annotations of the text entities are defined in an ontology that can be revised in parallel. AlvisAE also includes a tool for detection and resolution of annotation conflicts.The annotations can be stored in a database and queried. The annotations are entities, n-ary relations and groups. The entities can be discontinuous and overlapping. AlvisAE can take as input semantic pre-annotations automatically produced by NLP pipelines such as AlvisNLP/ML.

**BioYaTeA** [Golik et al., 2013] is an extension of the YaTeA term extractor [Aubin and Hamon, 2006] that deals with prepositional attachment and adjectival participles. It extracts terms from documents in French and in English. It includes post-filtering of irrelevant terms.

[3] http://informatique-mia.inra.fr/logiciels/node/272

[4] https://www.mathinfo.inra.fr/fr/content/knowledge-engineering-mathematics-and-informatics/alvisir-alvis-information-retrieval

[5] http://informatique-mia.inra.fr/logiciels/node/2

[6] http://informatique-mia.inra.fr/logiciels/node/273

[7] https://www.mathinfo.inra.fr/en/content/knowledge-engineering-mathematics-and-informatics/alvisae-alvis-annotation-editor

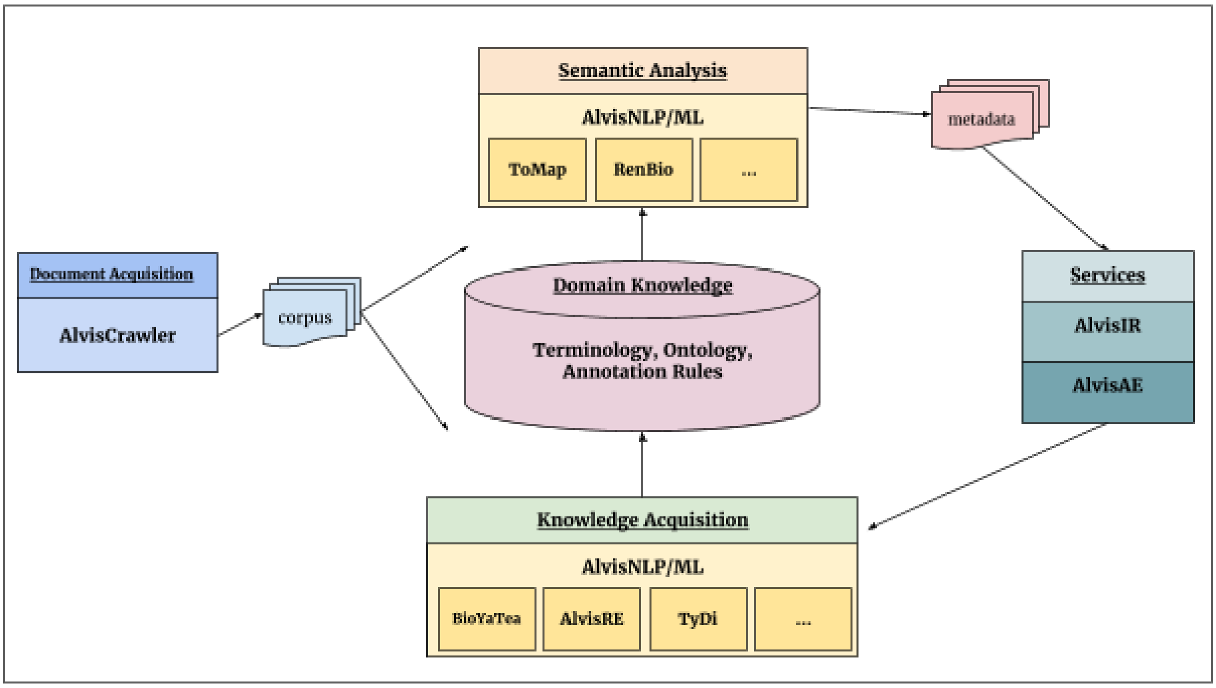


Figure 1.7: The architecture of Alvis Suite and its components.

**RenBio** is a program to identify gene and protein names in a textual document based on machine learning techniques. It searches for named entities in a document according to a decision tree, based on attributes such as regex matches, dictionary matches, *etc.*

**ToMap** (*Text to Ontology Mapping*)[Golik et al., 2011, Bossy et al., 2015] is a method that extracts and categorizes terms based on an ontology. It is based on the phrase similarity syntactic analysis principle, such as used in MetaMap [Aronson and Lang, 2010]. It is applicable to all types of termino-ontologies and corpora in French and English.

**TyDI** (*Terminology Design Interface*) [Nédellec et al., 2010] is a collaborative tool for the manual validation/annotation and structuring of terms either originating from terminologies or extracted from a training corpus of textual documents. It is used on the output of term extractors (like BioYatea). With TyDI, a user can validate candidate terms and specify synonymy/hyperonymy relations. These annotations can then be exported in several formats, and used in other natural language processing tools (such as AlvisNLP/ML). It shares its ontology with AlvisAE so that collective ontology revisions can be managed through both tools in real-time.

More information on Alvis Suite and its components can be found in Claire Nédellec's habilitation thesis [Nédellec, 2013] (in French).

## 1.8 Conclusion

*Arabidopsis thaliana* is a compelling subject for information extraction. As a model plant it serves as a prototype for biologists and any result on it can hope to be generalized to other organisms. *A. thaliana* and its seeds in particular are of interest for agriculture, further underlining its importance as an organism. Seed development and its regulatory networks are being heavily researched, but a need for augmenting the efforts and testing new sources of information has been identified by prominent experts of the field.

When building a model for knowledge extraction a number of often conflicting constraints have an effect on the procedure. First and foremost maximizing the scope and detail covered by such an extraction is an implied constraint. On the other hand, the tools used and their practical implications, such as the usability of interfaces call for simplicity. Similarly, linguistic and machine learning constraints call for easy generalization and sufficient volumes of manually annotated data. All these aspects need to be taken into consideration during model design and make it necessary to bring together multidisciplinary teams of experts in order to do so.

Two different models can be used in order to address the conflict between relevant simplicity needed by the annotation and prediction methods, on one hand, and the detail necessary to obtain satisfactory knowledge models. The annotation model is used during the knowledge extraction process and the knowledge model can be used later for direct use or other applications. A correspondence between these models needs naturally to be defined during the model design phase.

Information Extraction has been attracting researchers for three decades and the methods proposed have evolved since its early years. These systems have evolved in both the information they take into account and their prediction algorithms. For Relation Extraction, the direction which the state of the art follows is using a maximum of linguistic information (surface, syntax, semantics) and Machine Learning. As far as learning is concerned, one can observe a shift toward the use of kernel methods which avoid the tasks of feature engineering and selection. Syntax graphs augmented by other linguistic information provide a good basis for a competitive kernel-based RE system.

Corpus annotation is a tedious and time consuming task which calls for the participation of experts of the target knowledge domain. This inherent cost of specialized corpora is the reason behind their rarity. Producing such a corpus is extremely useful for both the target domain and the IE community and needs to be done according to established methods and norms. In order to benefit the communities, it is a basic requirement that any corpus produced comes with a good documentation and public availability.

# Chapter 2

# Data

**2.1 Introduction** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **38**
2.1.1 Tasks & Development Phases . . . . . . . . . . . . . . . . . . . . . . . . . 38
2.1.2 People . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39
**2.2 Tools** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **40**
2.2.1 Collaborative Tools . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
2.2.2 Documentation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
2.2.3 Annotation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43
Annotation Editor . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43
Adjudication . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43
**2.3 Model** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **43**
2.3.1 Conceptual Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48
Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48
Entities . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48
Relations and Events . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50
2.3.2 Annotation Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55
Entities . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
Relations and Events . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
2.3.3 Model Transformations . . . . . . . . . . . . . . . . . . . . . . . . . . . 60
Transformation for Relation Extraction . . . . . . . . . . . . . . . . . . 60
**2.4 Corpus** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **64**
2.4.1 Source . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 64
2.4.2 Corpus Annotation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 68
Automatic pre-annotation . . . . . . . . . . . . . . . . . . . . . . . . . . 68
Campaigns . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 68
Adjudication . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 69
**2.5 Results and Discussion** . . . . . . . . . . . . . . . . . . . . . . . . . . . . **70**
**2.6 Conclusion** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **74**

## 2.1 Introduction

The first research question adressed in this work seeks to find a compromise between describing adequately the target biological domain, while at the same time producing a corpus that is adapted to machine learning. This chapter is dedicated to the model and corpus on seed development for *A. Thaliana*, and the focus of this work has been high quality, representativeness, and reusability for both the model and the corpus.

This project is interdisciplinary, but it was also a complex and complete one: it went from the conception of the model to the actual text-mining application. The two main consequences of the nature of this project are the necessity of the collaboration between a number of people from different backgrounds who did not know each other before, and a big investment in time in order to complete the various necessary tasks.

Seed development for *A. Thaliana* has been the main focus of my work. Nevertheless, it is noteworthy to mention that I also participated in the production of the Bacteria Biotopes corpus for the BioNLP challenge – a work presented in detail in [Bossy et al., 2015], which I also used as an intermediary step for the verification of my algorithmic approach.

### 2.1.1 Tasks & Development Phases

A corpus project such as this one delivers at least three distinct products: the model, the documentation and, finally, the annotated corpus. Before expanding on the composition of the team working on this corpus, I would like to present the development phases, in chronological order.

**Phase 1: Model Design**

This initial phase served to conceptualize the model that was going to be used in the following steps. It defined the scope of the knowledge to cover, took into account the needs of the various parties and compromises these called for and defined the entities and relations. Additionally, the documentation started to be written during this phase.

**Phase 2: Iterative Model Development**

Once the model had been initialized, an iterative development phase started, by confronting the model and the text. This process was coupled with the annotation phase, as it is the latter that puts the model into use by the experts and provides the necessary feedback for model improvement. During this step the model was studied and revised example by example, resulting in precisions, expansions and re-definitions of entities and relations. The model adjustments include additions, splits, merges and removal of types resulting from an iterative exchange process between the team members, via communication channels such as meetings, a dedicated forum and messages. The annotation guidelines document was consequently modified and enriched accordingly, with examples but also counter-examples.

**Phase 3: Annotation**

Corpus annotation was organized in campaigns. Apart from the annotators, the annotation campaigns called for coordination and follow up during the annotation process, as well as the adjudication and finalization steps.

#### Phase 4: Model Transformations

The initial model was designed for manual annotation, taking into account constraints such as user interface and ease of use. After the annotation was completed, a rewriting of the model took place. This transformed version of the model is more focused on being conceptually correct and usable by other applications. A special transformed version of the model and data was prepared for the BioNLP 2016 Shared Task SeeDev[8], for the subtask called "seedev-full". Finally, a set of rules for binary rewriting of multi-entity events was also carefully studied and produced for machine-learning purposes, which was also used for the "seedev-binary" subtask of BioNLP-ST.

### 2.1.2 People

For this part of the project to succeed tools and expertise on multiple different axes were needed. First of all, it was necessary to collaborate with biologists and experts on *A. Thaliana* and its seed development, in particular. Knowledge modeling was necessary for the conception part, as was annotation software for the annotation phase. Finally, we thought it was important to involve researchers with perspective on similar projects in the biomedical domain.

A list of the collaborators in this project, in alphabetical order, presenting their profile and summarizing their contribution can be found below:

1. **Philippe Bessières** is a senior scientist, expert in biology with a long experience in information extraction, knowledge modelling and corpus production, especially for biology. His contributions were mainly consultative on all phases, providing perspective from a biological point of view and experience with other IE corpora production.

2. **Robert Bossy** is a senior research engineer, with expertise in information extraction and a background in biology. He developed software tools whenever necessary, namely contributing to the annotation editor (AlvisAE [Papazian et al., 2012]), to the automatic pre-annotation by the Alvis Suite [Ratkovic et al., 2011, Nédellec et al., 2009] and to the model transformation. He also played a consultative role throughout the project with regards to both IE and biology.

3. **Estelle Chaix** is a post-doctoral researcher with a biology background. She was responsible for the coordination of the annotation campaigns, by managing the communication and schedule, annotating, correcting and coordinating the adjudication. She was also responsible for the feedback from the annotation campaign and the adjustments to the model, as well as the transformations (phases 2, 4). Finally, Estelle gave presentations to the Bio and IE communities (talks, poster [Chaix et al., 2015a, Chaix, 2015]).

[8] http://2016.bionlp-st.org/tasks/seedev

4. **Louise Deléger** is a research scientist, expert on natural language processing, information extraction, knowledge modelling and corpus production, particularly in the biomedical domain. She contributed by consulting in phases 2-4 and participating in the adjudication and model transformation steps.

5. **Bertrand Dubreucq** is a senior scientist, expert on *A. Thaliana* and seed development in particular. He contributed to model design and more generally provided perspective from a biological point of view (phases 1-4). He was one of the main expert annotators. Finally, he presented this project to the *Arabidopsis* and bioinformatics communities (talks, poster [Dubreucq et al., 2015, Dubreucq, 2014, Nédellec and Dubreucq, 2014]).

6. **Abdelhak (Abdou) Fatihi** is a post-doctoral researcher with a biology background, specialized in *A. thaliana* and seed development in general. He was one of the main annotators and also contributed to model development.

7. **Loïc Lepiniec** is a senior scientist, expert on *A. Thaliana* and seed development in particular. He contributed to model design and biology perspective throughout the project and he equally contributed to the first annotation campaign.

8. **Claire Nédellec** is a senior scientist, expert on information extraction, knowledge modelling and corpus production, particularly in the domain of biology. She coordinated this project and provided major contributions in the model design and development (phases 1, 2, 4).

9. **Frédéric Papazian** is the software engineer who was responsible for the development of the annotation editor used for this project, AlvisAE. He developed new utilities to facilitate manual annotation and adjudication and made user interface adjustments whenever necessary.

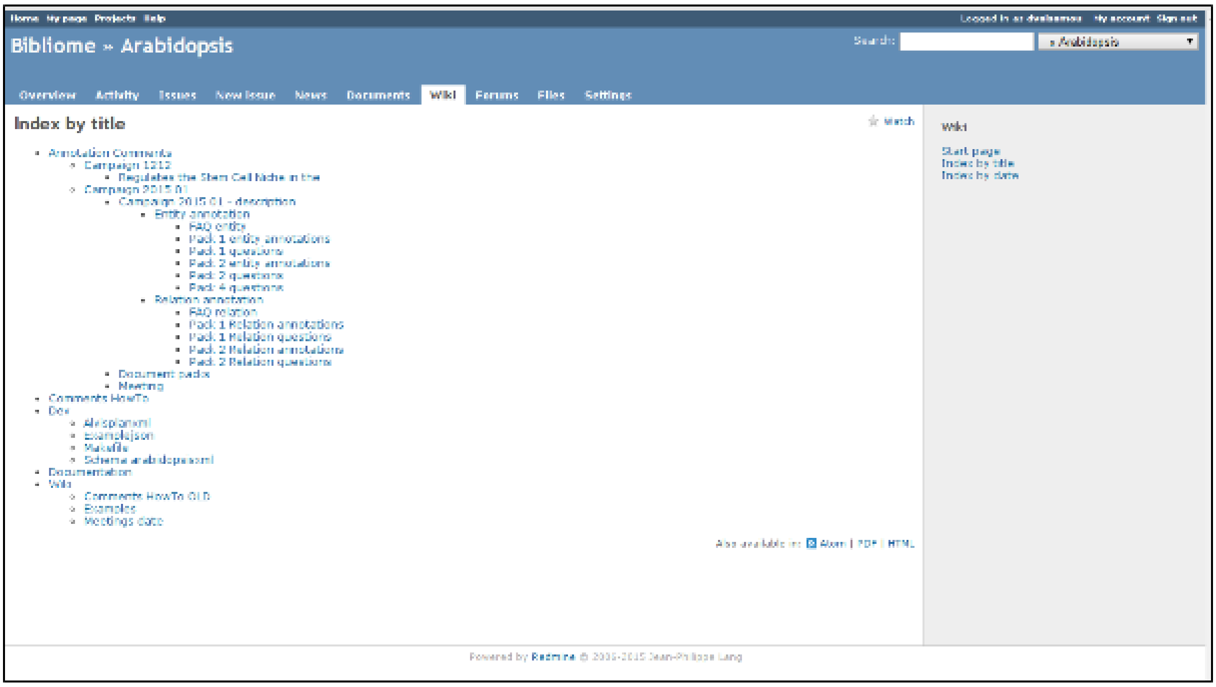
10. **Pierre Zweigenbaum** is a senior scientist, expert on information extraction, knowledge modelling and corpus production, particularly in the medical domain, as well as natural language processing. He played a consultative role on model design and development (phases 1, 2, 4) and provided perspective for the information extraction aspects of the project.

11. Finally, I personally contributed mainly during phase 1, working on model design and the first complete versions of the guidelines document [Chaix et al., 2015b]. I continued contributing during phases 2-4, particularly in the first annotation campaigns and revisions of the model and the documentation, as well as by presenting the project (*e.g.* [Valsamou, 2013, Valsamou and Nedellec, 2012, Valsamou, 2015]).

## 2.2 Tools

### 2.2.1 Collaborative Tools

For the organization of the annotation process the project management tool Redmine [Lang and Davis, 2010] was used. Redmine is a web-based tool, providing a dedicated website



Index by title

- Annotation Comments
  - Campaign 1212
    - Regulates the Stem Cell Niche in the
  - Campaign 2015 01
    - Campaign 2015 01 - description
      - Entity annotation
        - FAQ entity
        - Pack 1 entity annotations
        - Pack 1 questions
        - Pack 2 entity annotations
        - Pack 2 questions
        - Pack 4 questions
      - Relation annotation
        - FAQ relation
        - Pack 1 Relation annotations
        - Pack 1 Relation questions
        - Pack 2 Relation annotations
        - Pack 2 Relation questions
    - Document packs
    - Meeting
- Comments HowTo
- Dev
  - Alvisplanxml
  - Examplejson
  - Makefile
  - Schema arabidopsisxml
- Documentation
- Wiki
  - Comments HowTo OLD
  - Examples
  - Meetings date

Figure 2.1: Redmine: the platform hosting the annotation website. This screenshot captures the index of pages.

for each project. The goal during the annotation phase was to have all relevant information available on the Redmine project website either as a dedicated page or an external link.

We used the website's wiki functions to create tutorials and centralize practical information. Wiki pages also served as a means for communication, as all members of the annotation team contributed examples, commentary, questions and answers.

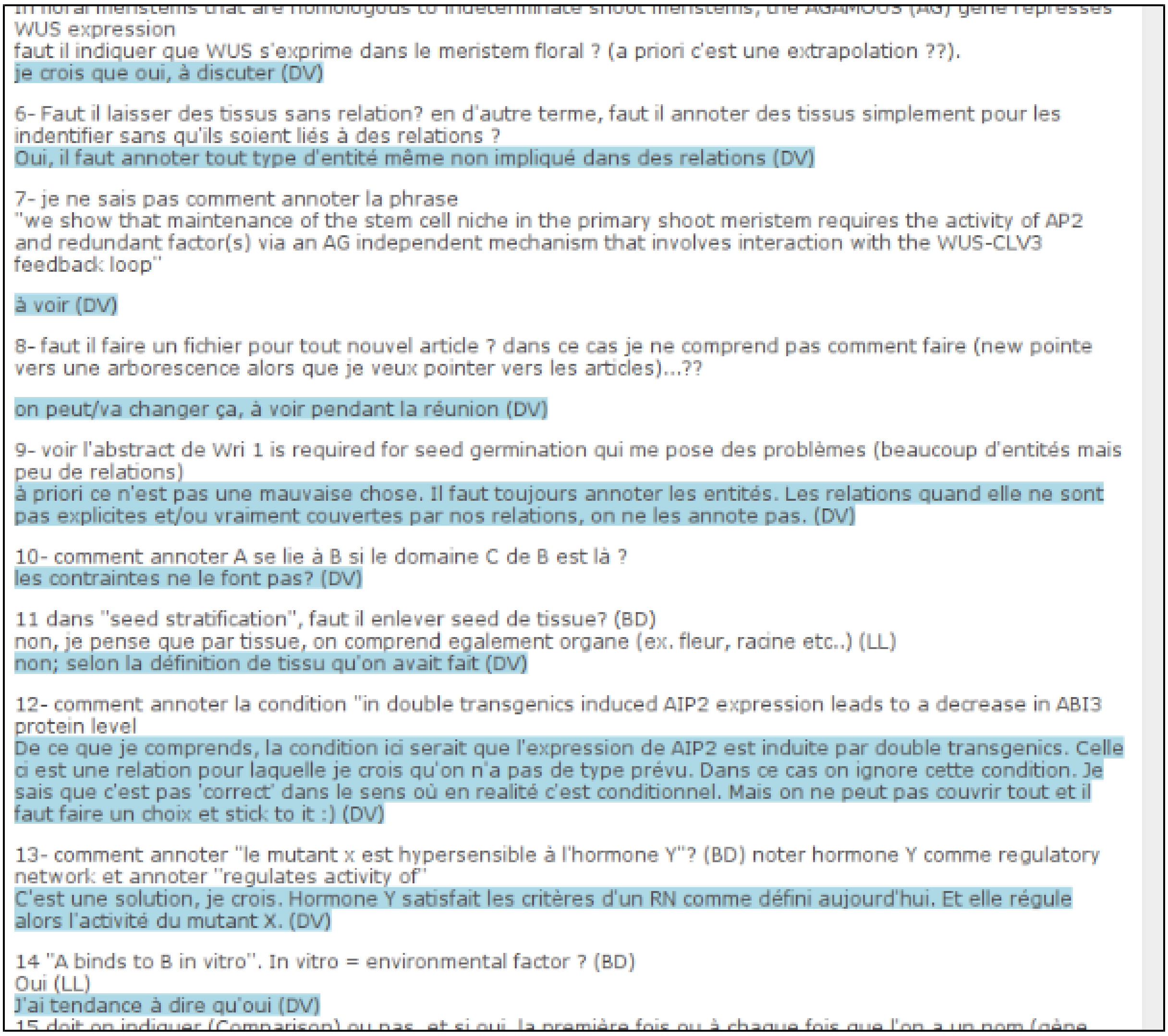
Finally, meetings took place on a bi monthly basis on average. Whereas periodical physical meetings with the majority of the team present were necessary for major revisions, regular follow-up and clarifications were assured by conference calls and reduced committee meetings.

### 2.2.2 Documentation

The central piece of documentation that aggregates all necessary information for the annotation model and process is the Guidelines document. In order to be able to work on the document collaboratively, we chose Google Docs[9] as a platform.

This document contains the model definitions, an extensive list of examples, counter-examples and clarifications and practical information for the annotators. In addition to the revision history capabilities Google Docs offer, a tracking system was used throughout the project, with major changes being listed on the first page of the document, for easier navigation.

[9] https://www.google.com/docs/about/



In floral meristems that are homologous to indeterminate shoot meristems, the AGAMOUS (AG) gene represses WUS expression
faut il indiquer que WUS s'exprime dans le meristem floral ? (a priori c'est une extrapolation ??).
je crois que oui, à discuter (DV)

6- Faut il laisser des tissus sans relation? en d'autre terme, faut il annoter des tissus simplement pour les indentifier sans qu'ils soient liés à des relations ?
Oui, il faut annoter tout type d'entité même non impliqué dans des relations (DV)

7- je ne sais pas comment annoter la phrase
"we show that maintenance of the stem cell niche in the primary shoot meristem requires the activity of AP2 and redundant factor(s) via an AG independent mechanism that involves interaction with the WUS-CLV3 feedback loop"

à voir (DV)

8- faut il faire un fichier pour tout nouvel article ? dans ce cas je ne comprend pas comment faire (new pointe vers une arborescence alors que je veux pointer vers les articles)...??

on peut/va changer ça, à voir pendant la réunion (DV)

9- voir l'abstract de Wri 1 is required for seed germination qui me pose des problèmes (beaucoup d'entités mais peu de relations)
à priori ce n'est pas une mauvaise chose. Il faut toujours annoter les entités. Les relations quand elle ne sont pas explicites et/ou vraiment couvertes par nos relations, on ne les annote pas. (DV)

10- comment annoter A se lie à B si le domaine C de B est là ?
les contraintes ne le font pas? (DV)

11 dans "seed stratification", faut il enlever seed de tissue? (BD)
non, je pense que par tissue, on comprend egalement organe (ex. fleur, racine etc..) (LL)
non; selon la définition de tissu qu'on avait fait (DV)

12- comment annoter la condition "in double transgenics induced AIP2 expression leads to a decrease in ABI3 protein level
De ce que je comprends, la condition ici serait que l'expression de AIP2 est induite par double transgenics. Celle ci est une relation pour laquelle je crois qu'on n'a pas de type prévu. Dans ce cas on ignore cette condition. Je sais que c'est pas 'correct' dans le sens où en realité c'est conditionnel. Mais on ne peut pas couvrir tout et il faut faire un choix et stick to it :) (DV)

13- comment annoter "le mutant x est hypersensible à l'hormone Y"? (BD) noter hormone Y comme regulatory network et annoter "regulates activity of"
C'est une solution, je crois. Hormone Y satisfait les critères d'un RN comme défini aujourd'hui. Et elle régule alors l'activité du mutant X. (DV)

14 "A binds to B in vitro". In vitro = environmental factor ? (BD)
Oui (LL)
J'ai tendance à dire qu'oui (DV)
15 doit on indiquer (Comparison) ou pas, et si oui, la première fois ou à chaque fois que l'on a un nom (gène

Figure 2.2: The wiki was used as a collaborative tool for sharing comments and questions.

### 2.2.3 Annotation

#### Annotation Editor

For the manual annotation by the experts, the online annotation editor AlvisAE [Papazian et al., 2012] was used. AlvisAE is a web browser-based annotation editor but also a framework which supports automatic annotations (linguistic or semantic), semantic relations and events as well as linking to an ontology. Its architecture as a client/server system allows for collaborative annotation, and its design as a browser-based application makes it portable and compatible with virtually any workstation. Figure 2.3 shows the main annotation view of AlvisAE.

AlvisAE supports an annotation model compatible with knowledge extraction projects, by design. The models presented in this chapter are directly usable in AlvisAE.

Finally, AlvisAE uses workflows aiming to centralize all operations throughout an annotation campaign (Figure 2.4). A campaign defines the tasks to be achieved on a given corpus and the user roles. For instance in this project, the annotation of the entities has been successively done in six steps, automatic pre annotation, manual revision and completion by Estelle Chaix, double-blind revision by the biologists, adjudication and final check by Estelle Chaix. As a result, in addition to the typical per user annotation view of one document, AlvisAE provides a comparative mode for adjudication (see fig. 2.5).

#### Adjudication

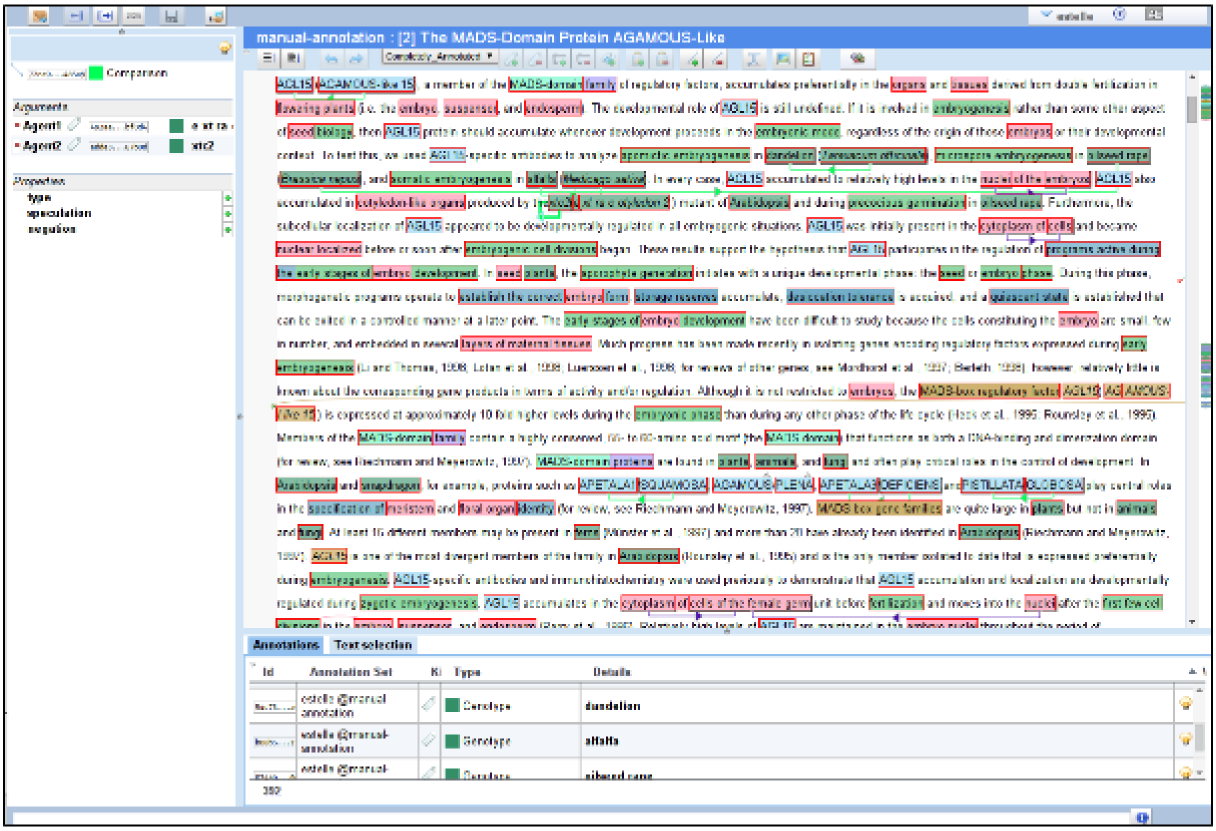
After the annotation phase, the annotations of the different annotators are compared and adjudicated in order to produce the *gold standard* that will be used for machine learning applications.

AlvisAE supports a side-to-side consolidation view of annotations for the adjudication phase (see Figure 2.5). While this mirrored view is very helpful because of its inline comparison, a second tool dedicated to adjudication was developed and used. This diff tool gives a detailed per document report, calculating statistical difference and inter-annotator agreement. Additionally, it provides a case per case analysis for the comparison of both entity and relation annotations between two annotators (Figures 2.7a-b).

## 2.3 Model

As mentioned before, there are multiple constraints that have to be satisfied by the model:

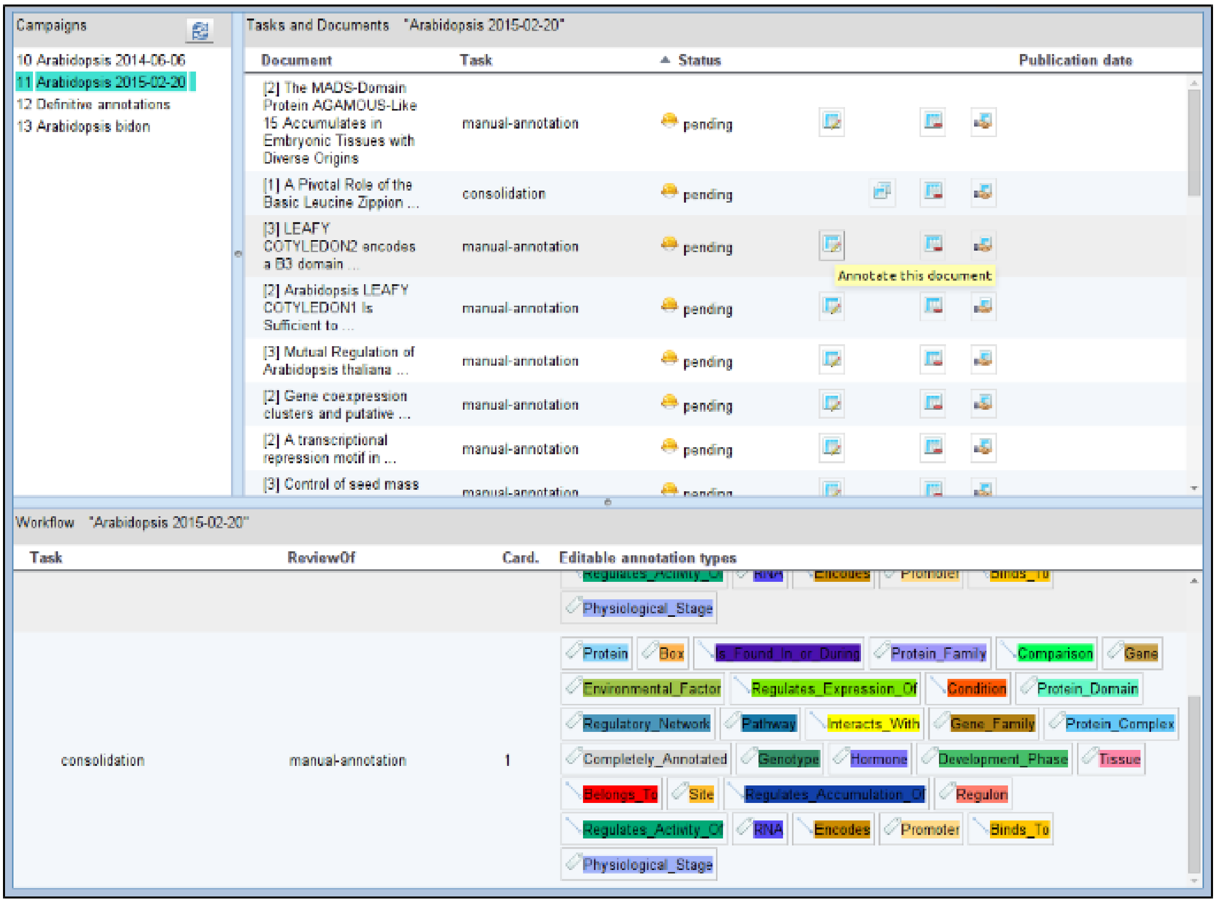
- the need to have a complete and precise conceptual model for the task of representing the knowledge of the biological domain,
- the need to be able to easily annotate the next manually with this model, and,
- the need to have a model that can be general enough for machine learning methods to adequately perform entity recognition and relation extraction.



AGL15 (AGAMOUS-like 15), a member of the MADS-domain family of regulatory factors, accumulates preferentially in the organs and tissues derived from double fertilization in flowering plants (i.e. the embryo, suspensor, and endosperm). The developmental role of AGL15 is still undefined. If it is involved in embryogenesis rather than some other aspect of seed biology, then AGL15 protein should accumulate whenever development proceeds in the embryonic mode, regardless of the origin of those embryos or their developmental context. To test this, we used AGL15-specific antibodies to analyze apomictic embryogenesis in dandelion (Taraxacum officinale), microspore embryogenesis in oilseed rape (Brassica napus), and somatic embryogenesis in alfalfa (Medicago sativa). In every case, AGL15 accumulated to relatively high levels in the nuclei of the embryos. AGL15 also accumulated in cotyledon-like organs produced by the lec2 (leafy cotyledon 2) mutant of Arabidopsis and during precocious germination in oilseed rape. Furthermore, the subcellular localization of AGL15 appeared to be developmentally regulated in all embryogenic situations. AGL15 was initially present in the cytoplasm of cells and became nuclear localized before or soon after embryogenic cell divisions began. These results support the hypothesis that AGL15 participates in the regulation of programs active during the early stages of embryo development. In seed plants, the sporophyte generation initiates with a unique developmental phase: the seed or embryo phase. During this phase, morphogenetic programs operate to establish the correct embryo form, storage reserves accumulate, desiccation tolerance is acquired, and a quiescent state is established that can be exited in a controlled manner at a later point. The early stages of embryo development have been difficult to study because the cells constituting the embryo are small, few in number, and embedded in several layers of maternal tissues. Much progress has been made recently in isolating genes encoding regulatory factors expressed during early embryogenesis (Li and Thomas, 1998; Lotan et al., 1998; Luerssen et al., 1998; for reviews of other genes, see Mordhorst et al., 1997; Berleth, 1998); however, relatively little is known about the corresponding gene products in terms of activity and/or regulation. Although it is not restricted to embryos, the MADS-box regulatory factor AGL15 (AGAMOUS-like 15) is expressed at approximately 10-fold higher levels during the embryonic phase than during any other phase of the life cycle (Heck et al., 1995; Rounsley et al., 1995). Members of the MADS-domain family contain a highly conserved 56- to 60-amino acid motif (the MADS domain) that functions as both a DNA-binding and dimerization domain (for review, see Riechmann and Meyerowitz, 1997). MADS-domain proteins are found in plants, animals, and fungi and often play critical roles in the control of development. In Arabidopsis and snapdragon, for example, proteins such as APETALA1/SQUAMOSA, AGAMOUS/PLENA, APETALA3/DEFICIENS and PISTILLATA/GLOBOSA play central roles in the specification of meristem and floral organ identity (for review, see Riechmann and Meyerowitz, 1997). MADS box gene families are quite large in plants but not in animals and fungi. At least 15 different members may be present in ferns (Münster et al., 1997) and more than 20 have already been identified in Arabidopsis (Riechmann and Meyerowitz, 1997). AGL15 is one of the most divergent members of the family in Arabidopsis (Rounsley et al., 1995) and is the only member isolated to date that is expressed preferentially during embryogenesis. AGL15-specific antibodies and immunohistochemistry were used previously to demonstrate that AGL15 accumulation and localization are developmentally regulated during zygotic embryogenesis. AGL15 accumulates in the cytoplasm of cells of the female germ unit before fertilization and moves into the nuclei after the first few cell divisions in the embryo, suspensor, and endosperm (Perry et al., 1996). Relatively high levels of AGL15 are maintained in the embryo nuclei throughout the period of

| Id | Annotation Set | Ki | Type | Details |
|---|---|---|---|---|
| [illegible] | estelle @manual-annotation |  | 🟩 Genotype | dandelion |
| [illegible] | estelle @manual-annotation |  | 🟩 Genotype | alfalfa |
| [illegible] | estelle @manual-annotation |  | 🟩 Genotype | oilseed rape |

Figure 2.3: AlvisAE, the annotation editor. A document from pack 2 is being annotated. The selected relation is a Comparison.



Campaigns

- 10 Arabidopsis 2014-06-06
- 11 Arabidopsis 2015-02-20
- 12 Definitive annotations
- 13 Arabidopsis bidon

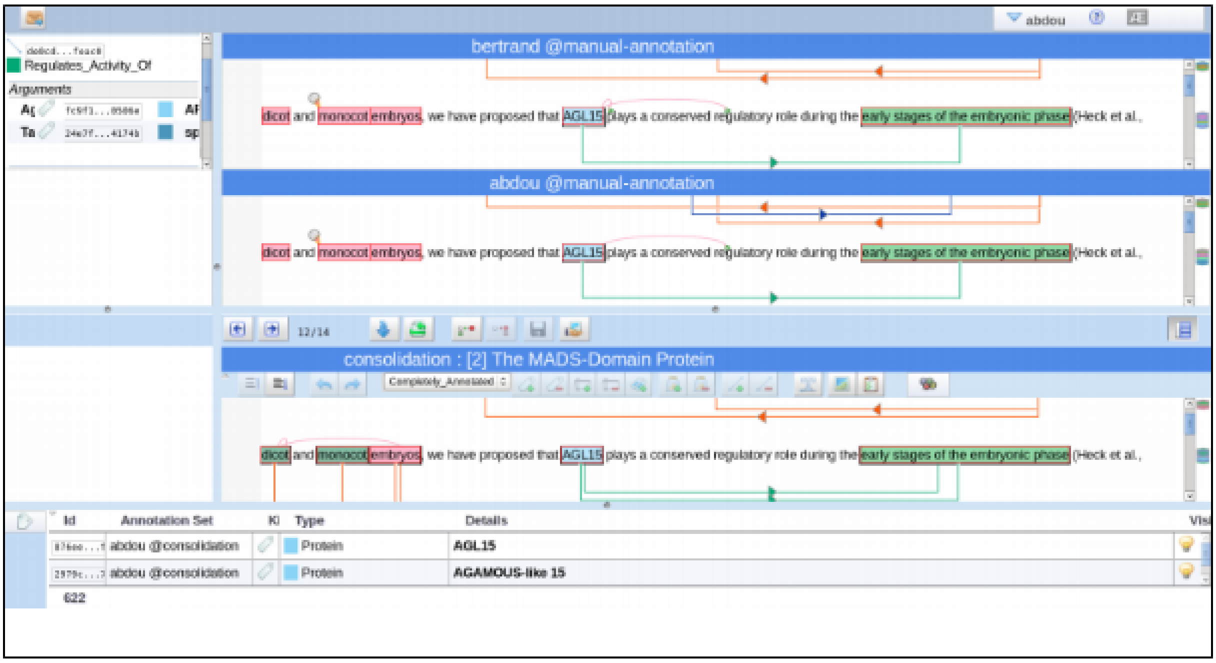
Tasks and Documents "Arabidopsis 2015-02-20"

| Document | Task | Status | Publication date |
|---|---|---|---|
| [2] The MADS-Domain Protein AGAMOUS-Like 15 Accumulates in Embryonic Tissues with Diverse Origins | manual-annotation | pending | |
| [1] A Pivotal Role of the Basic Leucine Zippion ... | consolidation | pending | |
| [3] LEAFY COTYLEDON2 encodes a B3 domain ... | manual-annotation | pending | |
| [2] Arabidopsis LEAFY COTYLEDON1 Is Sufficient to ... | manual-annotation | pending | |
| [3] Mutual Regulation of Arabidopsis thaliana ... | manual-annotation | pending | |
| [2] Gene coexpression clusters and putative ... | manual-annotation | pending | |
| [2] A transcriptional repression motif in ... | manual-annotation | pending | |
| [3] Control of seed mass | manual-annotation | pending | |

Workflow "Arabidopsis 2015-02-20"

| Task | ReviewOf | Card. | Editable annotation types |
|---|---|---|---|
| | | | Regulates_Activity_Of RNA Encodes Promoter Binds_To<br>Physiological_Stage |
| consolidation | manual-annotation | 1 | Protein Box Is_Found_In_or_During Protein_Family Comparison Gene<br>Environmental_Factor Regulates_Expression_Of Condition Protein_Domain<br>Regulatory_Network Pathway Interacts_With Gene_Family Protein_Complex<br>Completely_Annotated Genotype Hormone Development_Phase Tissue<br>Belongs_To Site Regulates_Accumulation_Of Regulon<br>Regulates_Activity_Of RNA Encodes Promoter Binds_To<br>Physiological_Stage |

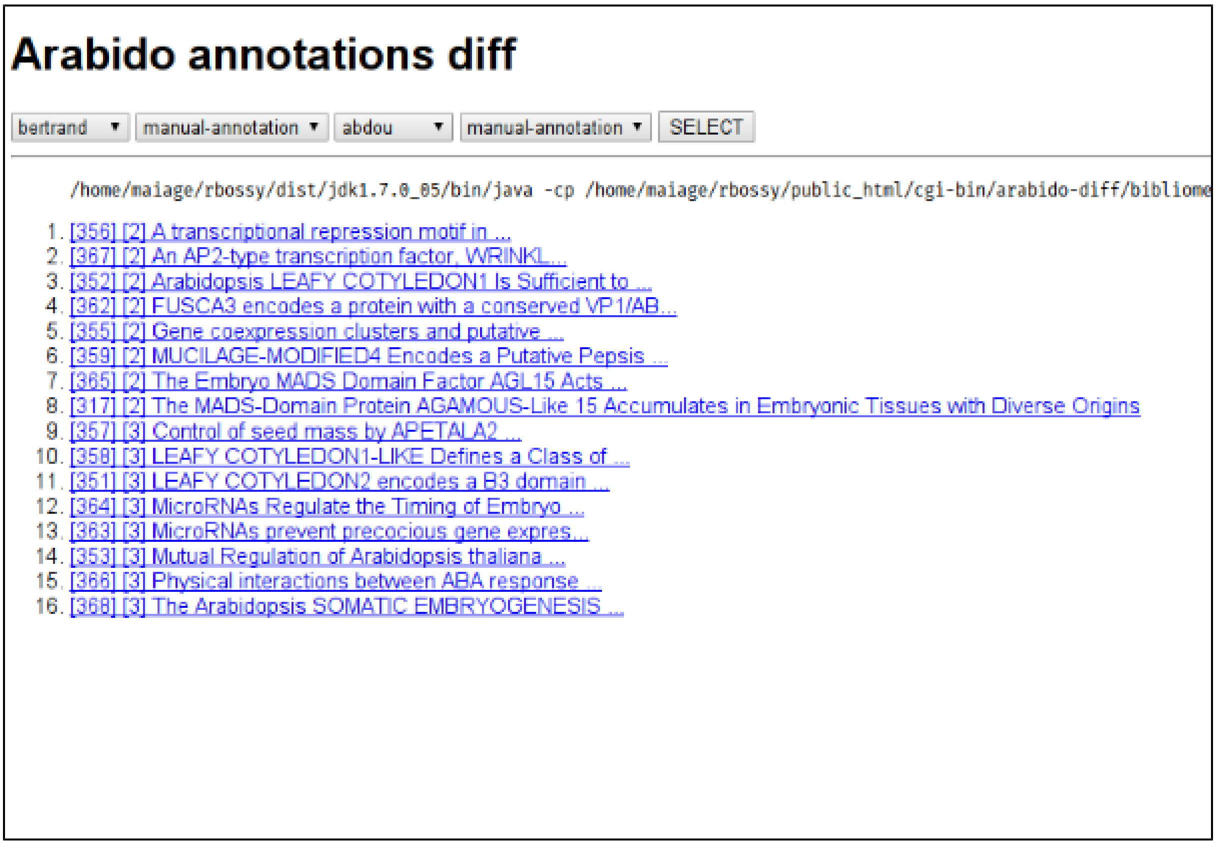
Figure 2.4: Campaign and workflow view in AlvisAE.



dicot and monocot embryos, we have proposed that AGL15 plays a conserved regulatory role during the early stages of the embryonic phase (Heck et al.,

| Id | Annotation Set | Kl | Type | Details | Visi |
|---|---|---|---|---|---|
| 876ee...1 | abdou @consolidation |  | Protein | AGL15 |  |
| 2979c...7 | abdou @consolidation |  | Protein | AGAMOUS-like 15 |  |
| 622 |  |  |  |  |  |

Figure 2.5: Consolidation view of a document from pack 2. The upper half of the window contains the biologist's annotations, while the bottom half contains the "gold standard" annotations.

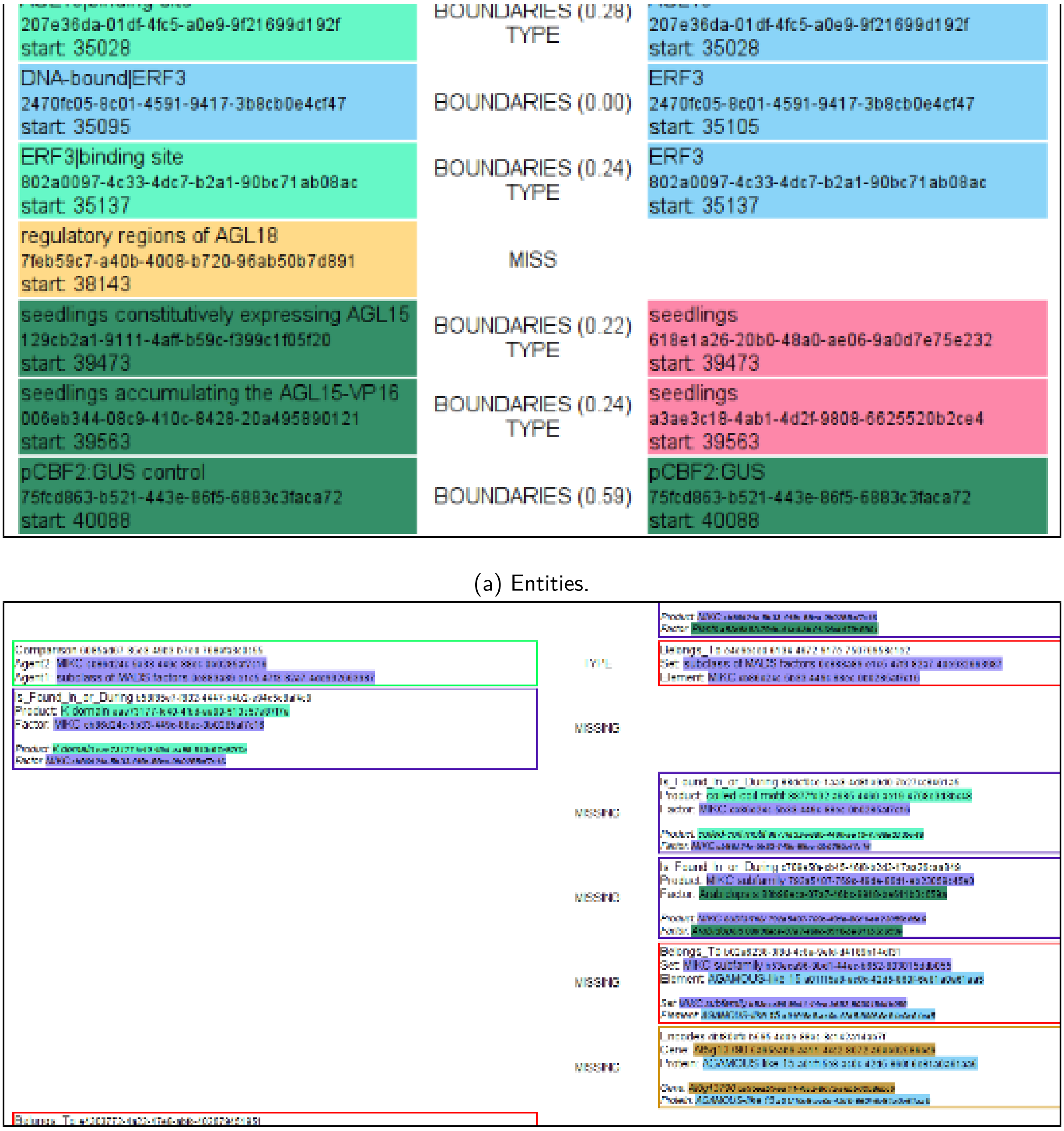

# Arabido annotations diff

bertrand ▼ manual-annotation ▼ abdou ▼ manual-annotation ▼ SELECT

```
/home/maiage/rbossy/dist/jdk1.7.0_05/bin/java -cp /home/maiage/rbossy/public_html/cgi-bin/arabido-diff/bibliome
```

1. [356] [2] A transcriptional repression motif in ...
2. [367] [2] An AP2-type transcription factor, WRINKL...
3. [352] [2] Arabidopsis LEAFY COTYLEDON1 Is Sufficient to ...
4. [362] [2] FUSCA3 encodes a protein with a conserved VP1/AB...
5. [355] [2] Gene coexpression clusters and putative ...
6. [359] [2] MUCILAGE-MODIFIED4 Encodes a Putative Pepsis ...
7. [365] [2] The Embryo MADS Domain Factor AGL15 Acts ...
8. [317] [2] The MADS-Domain Protein AGAMOUS-Like 15 Accumulates in Embryonic Tissues with Diverse Origins
9. [357] [3] Control of seed mass by APETALA2 ...
10. [358] [3] LEAFY COTYLEDON1-LIKE Defines a Class of ...
11. [351] [3] LEAFY COTYLEDON2 encodes a B3 domain ...
12. [364] [3] MicroRNAs Regulate the Timing of Embryo ...
13. [363] [3] MicroRNAs prevent precocious gene expres...
14. [353] [3] Mutual Regulation of Arabidopsis thaliana ...
15. [366] [3] Physical interactions between ABA response ...
16. [368] [3] The Arabidopsis SOMATIC EMBRYOGENESIS ...

Figure 2.6: The diff tool developed for the adjudication phase.



| | | |
|---|---|---|
| 207e36da-01df-4fc5-a0e9-9f21699d192f<br>start: 35028 | BOUNDARIES (0.28)<br>TYPE | 207e36da-01df-4fc5-a0e9-9f21699d192f<br>start: 35028 |
| DNA-bound\|ERF3<br>2470fc05-8c01-4591-9417-3b8cb0e4cf47<br>start: 35095 | BOUNDARIES (0.00) | ERF3<br>2470fc05-8c01-4591-9417-3b8cb0e4cf47<br>start: 35105 |
| ERF3\|binding site<br>802a0097-4c33-4dc7-b2a1-90bc71ab08ac<br>start: 35137 | BOUNDARIES (0.24)<br>TYPE | ERF3<br>802a0097-4c33-4dc7-b2a1-90bc71ab08ac<br>start: 35137 |
| regulatory regions of AGL18<br>7feb59c7-a40b-4008-b720-96ab50b7d891<br>start: 38143 | MISS | |
| seedlings constitutively expressing AGL15<br>129cb2a1-9111-4aff-b59c-f399c1f05f20<br>start: 39473 | BOUNDARIES (0.22)<br>TYPE | seedlings<br>618e1a26-20b0-48a0-ae06-9a0d7e75e232<br>start: 39473 |
| seedlings accumulating the AGL15-VP16<br>006eb344-08c9-410c-8428-20a495890121<br>start: 39563 | BOUNDARIES (0.24)<br>TYPE | seedlings<br>a3ae3c18-4ab1-4d2f-9808-6625520b2ce4<br>start: 39563 |
| pCBF2:GUS control<br>75fcd863-b521-443e-86f5-6883c3faca72<br>start: 40088 | BOUNDARIES (0.59) | pCBF2:GUS<br>75fcd863-b521-443e-86f5-6883c3faca72<br>start: 40088 |

(a) Entities.

| | | |
|---|---|---|
| | | Product: MIKC [illegible]<br>Factor: [illegible] |
| Comparison [illegible]<br>Agent2: MIKC [illegible]<br>Agent1: subclass of MADS factors [illegible] | TYPE | Belongs_To [illegible]<br>Set: subclass of MADS factors [illegible]<br>Element: MIKC [illegible] |
| Is_Found_In_or_During [illegible]<br>Product: K domain [illegible]<br>Factor: MIKC [illegible]<br>Product: K domain [illegible]<br>Factor: MIKC [illegible] | MISSING | |
| | MISSING | Is_Found_In_or_During [illegible]<br>Product: coiled coil motif [illegible]<br>Factor: MIKC [illegible]<br>Product: coiled coil motif [illegible]<br>Factor: MIKC [illegible] |
| | MISSING | Is_Found_In_or_During [illegible]<br>Product: MIKC subfamily [illegible]<br>Factor: Arabidopsis [illegible]<br>Product: MIKC subfamily [illegible]<br>Factor: Arabidopsis [illegible] |
| | MISSING | Belongs_To [illegible]<br>Set: MIKC subfamily [illegible]<br>Element: AGAMOUS-like 15 [illegible]<br>Set: MIKC subfamily [illegible]<br>Element: AGAMOUS-like 15 [illegible] |
| | MISSING | Encodes [illegible]<br>Gene: At5g13790 [illegible]<br>Protein: AGAMOUS-like 15 [illegible]<br>Gene: At5g13790 [illegible]<br>Protein: AGAMOUS-like 15 [illegible] |
| Belongs_To [illegible] | | |

(b) Relations

Figure 2.7: Entity (a) and relation (b) comparison in the diff tool.

In order to achieve all of these goals at the same time, we produced two different versions of the model: the conceptual model, which offers a finer granularity in concept specialization, and the annotation model, which offers ease of use for annotators. Additionally, we defined a hierarchy of concepts whenever this was necessary, in order to use the more specialized concepts for knowledge representation, and the more general ones for prediction. The transformation between these different versions and granularities of the model is guaranteed by a set of rules.

In summary, four versions of the model exist:

- The conceptual model, having 16 entity types and 21 n-ary event types.
- The annotation model, with 16 entity types, 10 binary relation types and 1 special relation type, "Condition", which helps represent n-ary events.
- The model used for "seedev-full" in BioNLP '16 ST challenge, which is a rewriting of the conceptual model.
- The model used for "seedev-binary" in the same challenge, where the "seedev-full" model was transformed to only include binary relations. In total, this model includes 22 binary relation types and 16 entity types.

### 2.3.1 Conceptual Model

#### Introduction

The conceptual model presented here is called Gene Regulation Network for *Arabidopsis* (GRNA). The purpose of this model is to best represent the structure and relations of the concepts in a way that is similar to the way that an ontology or other knowledge base is organized. By working with leading experts in the domain, we assured that this model corresponds well to the biological world and knowledge extracted according to it meets the requirements for use in modeling in systems biology and integration with knowledge from experimental data.

Complex multi-argument relations and semantic differentiation of fine granularity are necessary for meaningful and useful knowledge models. For this reason, they are included in the conceptual model, even though they tend to be tedious for manual annotation. Favouring semantic precision by avoiding highly inclusive types can lead to underrepresentation of these types in the corpus, which is problematic for automatic information extraction. However, the GRNA model is designed with such applications in mind and by providing a hierarchy of types, it gives the choice of modulating granularity if necessary.

This section presents briefly all the types of entities and relations of the conceptual model, providing a short description and examples for each. A longer list of examples and special cases can be found in the official Guidelines document for the BioNLP ST '16 SeeDev Task [Chaix et al., 2016b]. Additionally, a detailed tabular view of the valid combination of arguments for the relations can be found in [Chaix et al., 2016c].

#### Entities

In order to keep the length of this chapter shorter, this and the following sections will only include short definitions of the types. Appendix E contains a more detailed presentation: a description, a list of short, characteristic examples and an example taken from the corpus, as well as the links to the sources and databases tied to this definition wherever applicable.

#### Gene

A gene is a DNA sequence coding for a mRNA. In the *A. thaliana* literature, genes are always written in uppercase and italics.

#### Gene Family

A family of genes mentioned by their common function, including coding for a same *protein family* or their common ancestor.

#### Promoter

A *Promoter* is an upstream region of a *Gene* that binds the polymerase for *Gene* transcription. It can be designated as a regulatory region of a *Gene.*

#### RNA

*RNA* is a gene product.

#### Protein

A *Protein* is an RNA product. Proteins are always in plain letters and uppercase.

#### Protein Family

A family of proteins mentioned by their common biologic function or by their common ancestor.

#### Protein Complex

A *Protein Complex* is a group of *Proteins* that physically interact together.

#### Protein Domain

A *Protein Domain* is a protein sequence and structure that can evolve, function and exist independently.

#### Hormone

A *Hormone* is a molecule that influences physiology and development.

### Pathway

*Pathway* means here metabolic pathway, for instance synthesis or degradation. A *pathway* represents a group of *genes* or corresponding *products* that are involved in a same metabolic, physiological or developmental *pathway.*

### Regulatory Network

A regulatory network is defined as a set of *Products* and/or *DNA* that control the expression of a *gene*, a *pathway.* It can also be used as a term to describe processes and functionality involving several *genes.* In all of these cases, a *regulatory network* corresponds to a regulatory function. This includes signalling pathways.

### Genotype

This type covers both genotypes and species. A *genotype* is given part or the whole genetic information (genetic composition) expressed by an organism genome. In this case, for *Arabidopsis thaliana.* A species is defined in reference to the biological nomenclature.

### Tissue

This type groups cell, tissue and organ. A *Tissue* is an ensemble of cells, not necessarily identical, that together carry out a specific function. Organs are then formed by the functional grouping together of multiple tissues. The *Tissue* type includes organs, as well as entities on the intra-cellular level, such as the “nuclei of the embryos” for example.

### Development Phase

A growth stage. This includes identity (cotyledon identity), dormancy (bud dormancy), development (cotyledon development) and growth (etiolated growth).

### Environmental Factor

Environmental or experimental conditions. This means any factor, abiotic or biotic, that influences living organisms (e.g. temperature, light, “in vitro”).

Figure 2.8 presents groupings or categories of types that have been defined for expressing the constraints on relation argument types in a concise way.

### Relations and Events

Relations in this model are defined as having two mandatory, and optionally more, arguments. For this reason, the terms *relation* and *event* are both applicable in this occasion. Obligatory arguments are those which are essential to the context of the relation, such as agent and target of a regulation, the member and the group in composition and membership, *etc.* Additional arguments correspond to conditions to these primary relations. These additional arguments are optional in the definition types not because they are not necessary for the annotation of an

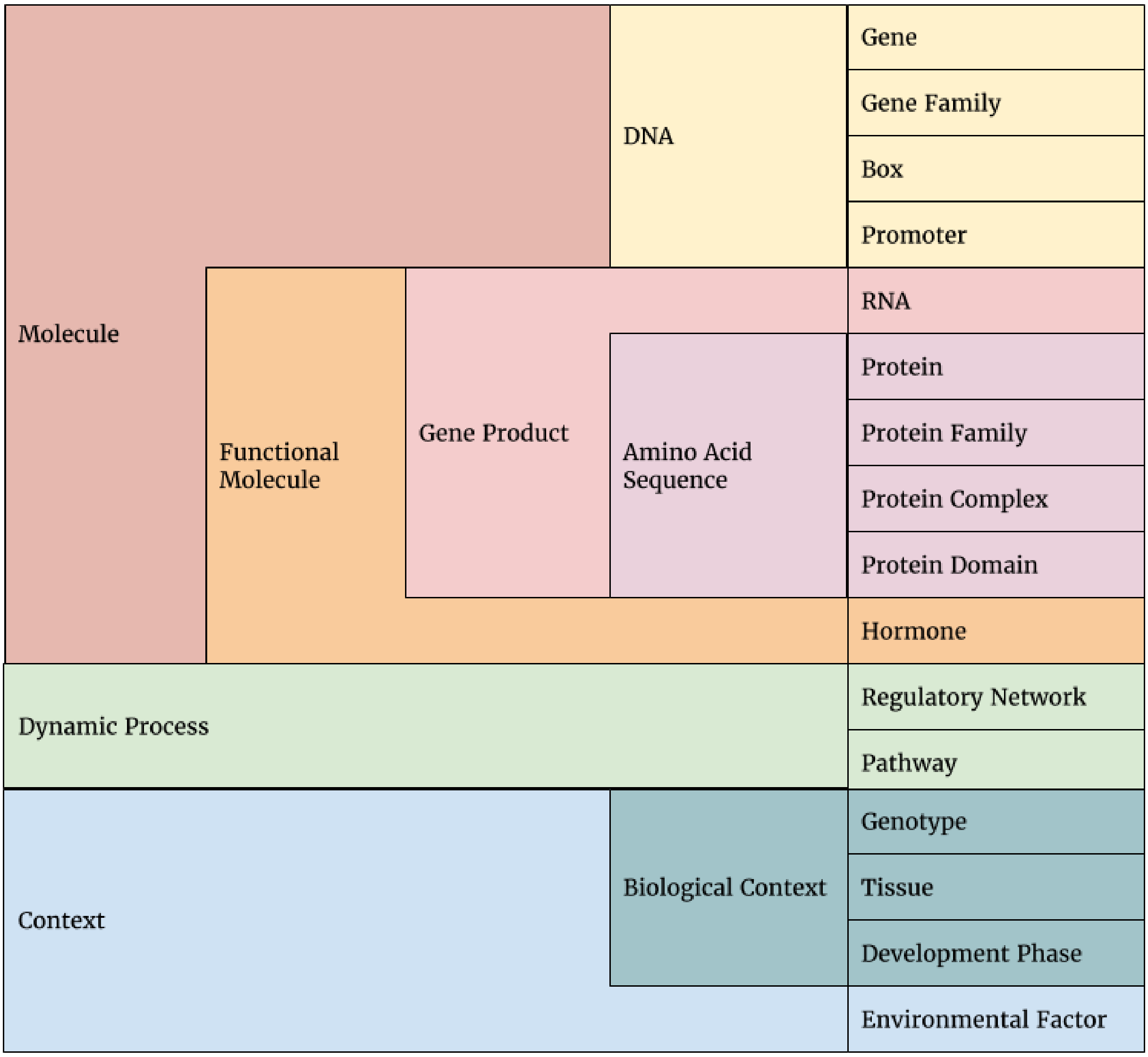


Figure 2.8: The hierarchy of entity types in the conceptual model

instance of a relation in which they take part, but because they are not always mentioned in the text. For example, if a regulation relation always occurs, only two arguments are necessary, but if it only occurs in a specific case –like a tissue or the pre-existence of another relation, then this relation instance will have a number of arguments corresponding to all necessary conditions mentioned. Consequently, these relation types are flexible in the number of arguments they can take –between 2 and 8 in the cases seen in this corpus.

The events of this model are n-ary and directed named (typed) relations between entity arguments belonging to the types defined above. All events have two required arguments and may have up to six secondary arguments. The roles of the arguments are also typed. Finally an event has two extra modifiers: negation, speculation.

As it was the case for the previous section, detailed descriptions are available in Appendix E. This section will simply list the types of relations along with a very short definition. For clarity, in the definitions below, argument names are written in **bold** letters, whereas entity or relation types are written in *italic.*

### Time and Localization

#### Presence In Genotype

A **Molecule** is present in a given **Genotype**.

#### Occurrence In Genotype

A (Dynamic) **Process** occurs in a given **Genotype**.

#### Presence At Stage

A *Functional Molecule* is present during a given *Development phase.*

#### Occurrence During

A **Process** occurs during a given *Development Phase.*

#### Localization

A *Functional Molecule* or *Dynamic Process* is found in a *Tissue*.

### Function

#### Involvement In Process

A *Molecule* is involved in a *Dynamic Process.*

### Transcription Or Translation

*DNA* entities encode for *RNA* (Transcription) or *RNA* entities encode *Proteins* (Translation). Often, reference is made to the *Gene* encoding the protein, without mention of the *RNA*.

### Functional Equivalence

A *Molecule*, *Dynamic Process* or *Context* compared to another similar *Molecule*, *Dynamic Process* or Context.

### Regulation

Regulation is one of the relation types that are defined in two granularities. For the machine learning prediction, one can choose to either grouping together all of the subtypes to have more examples, or to treat each subtype separately, as the argument combinations are unique for each subtype. And as far as knowledge representation is concerned, the subtypes correspond to related but different biological relations, so by defining these subtypes the model is more precise.

*Regulation* relation types are used when there is a *Genotype* involved. The **Agent** of this type of relation is always the *Genotype*, even when the *Gene* involved in the *Genotype* is specified, if no information on the direct role of the agent is given.

### Regulation Of Accumulation

An **Agent** regulates the accumulation of a *Molecule*.

### Regulation Of Expression

An **Agent** regulates the expression of a *DNA* entity.

### Regulation Of Development Phase

An **Agent** regulates the activity of a *Development Phase*.

### Regulation Of Molecule Activity

An **Agent** regulates the activity of a **Molecule**.

### Regulation Of Process

An **Agent** regulates the activity of a *Dynamic Process*.

### Regulation Of Tissue Development

An **Agent** regulates the development of a *Tissue*.

### Composition and Membership

### Primary Structure Composition

A specific sequence of nucleotide (**Box** or **Promoter**) is found in a molecule of **DNA.**

### Protein Complex Composition

An *Amino Acid Sequence* is found in a *Protein Complex***.**

### Protein Domain Composition

A specific *Protein Domain* is found in an *Amino Acid Sequence.*

### Family Membership

A *DNA***,** or *Gene Product* belongs to another *DNA***,** or *Gene Product.* This relation is to be used between entities of the same nature, to denote members of a set (e.g. *Gene* belonging to *Gene Family*, *Protein* to a *Protein Family*, sub-families to families, etc.).

### Sequence Identity

A *Molecule*, *Dynamic Process* or *Context* compared to another similar *Molecule*, *Dynamic Process* or *Context.* This type of relation is used for linking identical products, as well as synonyms, full form and abbreviation.

### Interaction

### Binding

A *Functional Molecule* physically binds to a *Molecule.* In most cases, a *Protein* binds to a *Promoter* or a *Gene.* An interaction between two proteins is specifically performed "*in vitro*" or in "yeast two-hybrid" and is annotated as a Binding relation.

### Interaction

A *Molecule* interacts with another *Molecule.* This type is used between DNA-DNA, in the case of indirect (non physical) interaction, and in any case of interaction where a more specific relation type cannot be used.

### Secondary Arguments

These optional secondary arguments are used to describe complex n-ary events and could serve the role of conditions or restraints. There are six types of secondary arguments, five for entities and one for events. Only one entity per role is possible for the entity secondary arguments, whereas multiple secondary events can be linked. All event types apart from *Presence in Genotype* accept secondary arguments.

1. **Tissue**: *Tissue*

2. **Development Stage**: *Development Phase*
3. **Organism Genotype**: *Genotype*
4. **Environmental Factor**: *Environmental Factor*
5. **Hormone**: *Hormone*
6. **Prerequisite Event:** *Primary Structure Composition | Interaction | Localization | Protein Domain Composition*

**Examples:**

- Entity argument: "A *Protein* accumulates in B *Tissue* when there exists C *Hormone*."
  - R1: *Localization* (**Functional Molecule**: A, **Target Tissue**: B, **Hormone** : C).
- Two entity arguments: "A *Protein* activates B *Gene* in the flower *Tissue*, if C *Hormone* increases."
  - R1: *Regulation Of Expression* (**Agent**: A, **DNA**: B , **Tissue**: flower , **Hormone**: C).
- Conjunction → two entity arguments: "A *Protein* accumulates in B *Tissue* when there exists C *Hormone and* D *Environmental Factor*."
  - R1: *Localization* (**Functional Molecule**: A, **Target Tissue**: B, **Hormone**: C, **Environmental Factor**: D).
- Disjunction → two unlinked relations: "A *Protein* accumulates in B *Tissue* when there exists C *Hormone or* D *Environmental Factor*."
  - R1: *Localization* (**Functional Molecule**: A, **Target Tissue**: B, **Hormone**: C).
  - R2 : *Localization* (**Functional Molecule**: A, **Target Tissue**: B, **Environmental Factor**: D).
- Relation argument: "A *Protein* binds to B *Protein complex* if C *Hormone* is found in D *Tissue*."
  - R1: *Localization* (**Functional Molecule**: B, **Target Tissue** D *Tissue*)
  - R2: *Binding* **(Agent**: A, **Target**: B, **Prerequisite Event**: R1)

### 2.3.2 Annotation Model

#### Introduction

This section covers the model as it was used for manual annotation. The goal of this model was to reduce the complexity of the model by finding the best compromise between a good representation of knowledge and ease of annotation. These compromises can be summarized by the following two points:

1. Binary relations are easier to annotate.

2. From a user interface and usability point of view, a tool to annotate multi-argument relations is tiresome to use and to visualize. On the other hand, binary relations can be used to connect all the arguments of one event. For this reason, we chose to annotate only two-argument relations, a choice that is often made in manual annotation campaigns (*e.g.* see the annotations for the GRO task of BioNLP'13[10]).

3. A complex model is harder for the annotator.

4. We tried to reduce the number of relation types whenever possible, by grouping together relationships which in this context are similar. We also defined subtypes of such more general relation types and added them as optional parameters in the model. This has two main advantages: a) the usability of the software, and b) the comprehension and memorization of the model by the annotators. Additionally, on the relation extraction step this choice of level of specificity *vs* generality can also be used to improve the learning rate.

In the following pages, a short description of the annotation model will be given. A more detailed explanation with numerous examples and clarifications can be found in the final version of the manual annotation campaign guidelines document [Chaix et al., 2015b]. For ease of comparison between the two models, names and examples remain similar whenever there is a correspondence.

### Entities

The entity types used for annotation are identical to these of the conceptual model (see section 2.3.1). The only change is a difference in the groups of entities, as they will be used for relation definitions. For the annotation model, the following groups are valid:

1. *DNA*: corresponding to the *DNA* group of the conceptual model.

2. *Product*: containing *Functional Molecule* and *Dynamic Process.*

3. *Factor*: which includes the *Context* entities of the conceptual model.

### Relations and Events

Annotated relations are binary, directed and named (typed) relations between entity arguments of the types defined above. The roles of the arguments of the relations are also typed.

**Conditions** are used to emulate n-ary events. These n-ary relations are represented by binary relations with one or more conditions. Conditions link binary relations and an argument. They are not typed.

Two types of **parameters** are used to represent qualifications of relations. The first type is modality, used for speculation and negation. As it was the case for the conceptual model, the annotation model will be presented in the section without detail, but Appendix F contains more detailed definitions and examples.

[10] http://pubannotation.org/projects/bionlp-st-gro-2013-training/docs

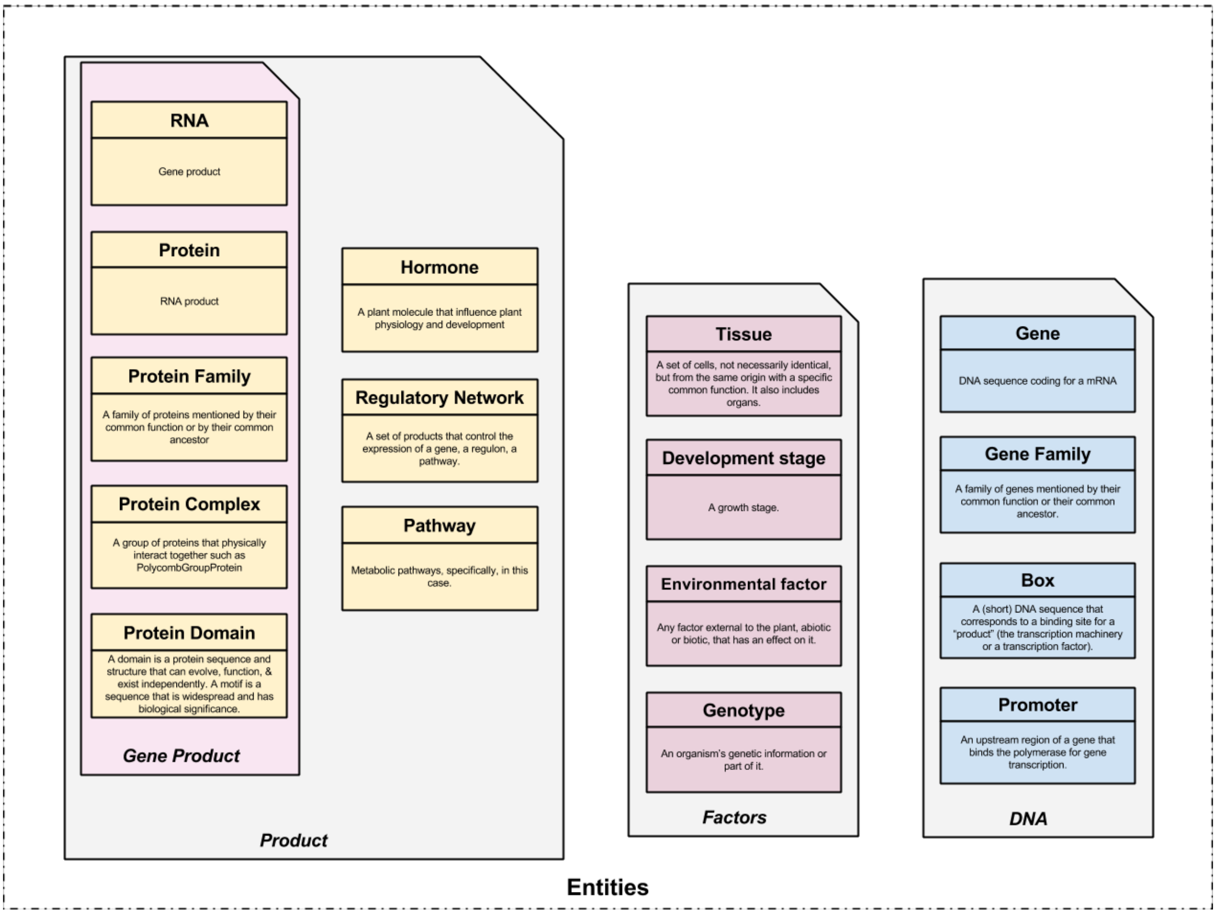

RNA
Gene product
Protein
RNA product
Protein Family
A family of proteins mentioned by their common function or by their common ancestor
Protein Complex
A group of proteins that physically interact together such as PolycombGroupProtein
Protein Domain
A domain is a protein sequence and structure that can evolve, function, & exist independently. A motif is a sequence that is widespread and has biological significance.
Gene Product
Hormone
A plant molecule that influence plant physiology and development
Regulatory Network
A set of products that control the expression of a gene, a regulon, a pathway.
Pathway
Metabolic pathways, specifically, in this case.
Product
Tissue
A set of cells, not necessarily identical, but from the same origin with a specific common function. It also includes organs.
Development stage
A growth stage.
Environmental factor
Any factor external to the plant, abiotic or biotic, that has an effect on it.
Genotype
An organism's genetic information or part of it.
Factors
Gene
DNA sequence coding for a mRNA
Gene Family
A family of genes mentioned by their common function or their common ancestor.
Box
A (short) DNA sequence that corresponds to a binding site for a "product" (the transcription machinery or a transcription factor).
Promoter
An upstream region of a gene that binds the polymerase for gene transcription.
DNA
Entities

Figure 2.9: The hierarchy of entity types in the annotation model

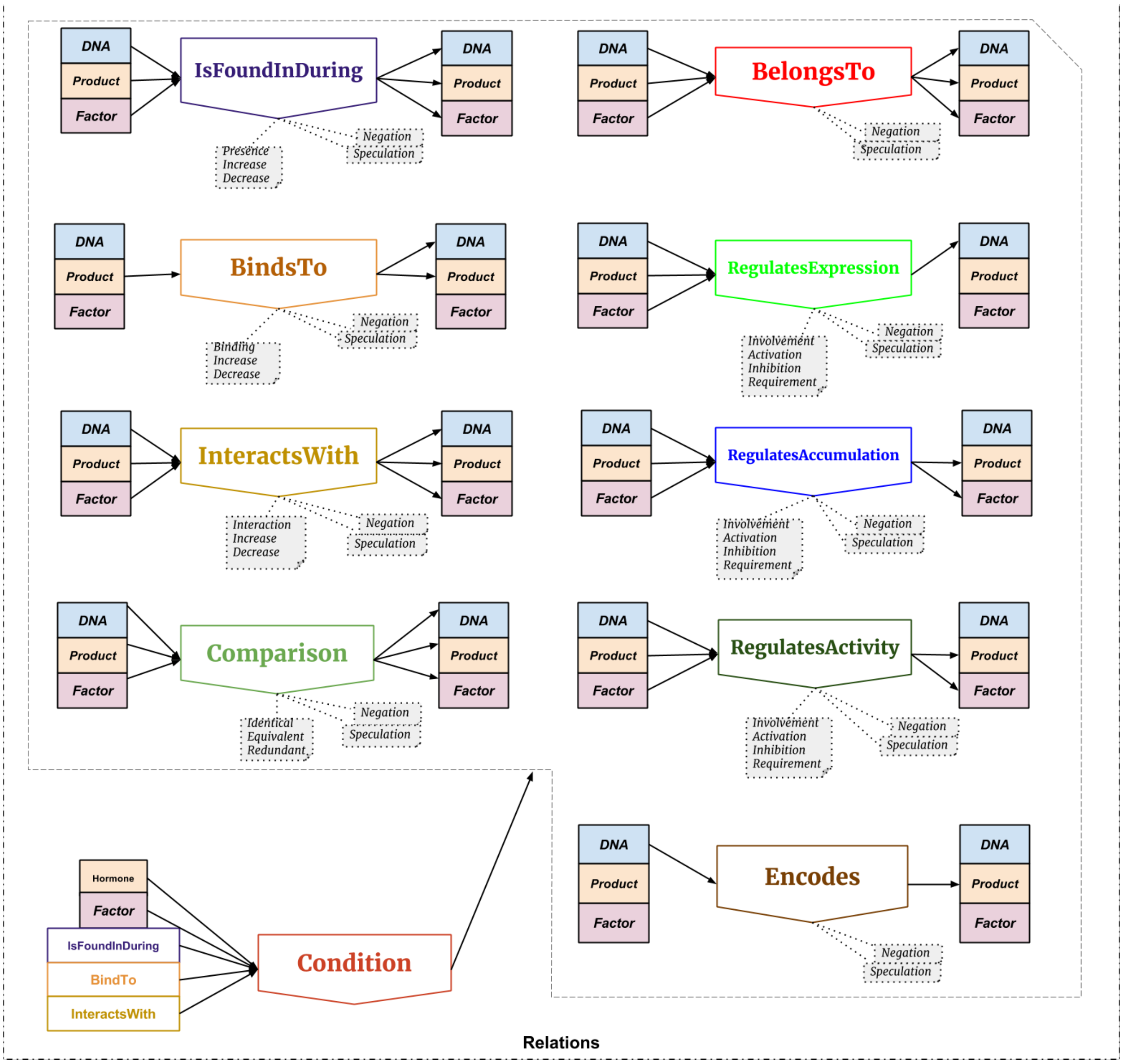


Figure 2.10: A schematic representation of relation types in the annotation model

### Interaction

#### Binds To

A *Product* physically binds to *DNA* or another *Product*. In most cases, a *Protein* binds to a *Promoter* or a *Gene*.

#### Interacts With

A *DNA*, *Product* or *Factor* interacts with another *DNA*, *Product*, or *Factor*, directly or indirectly. It is used whenever the more specialized *BindsTo* (for direct physical interactions), *Regulates Activity Of* and *Regulates Expression Of* are not appropriate.

### Similarity

This group corresponds to two groups in the conceptual model: Function, and Composition and Membership.

#### Encodes

A *DNA* or *RNA* entity encodes a *Product* such as a *Protein*.

#### Belongs To

A *DNA, Product* or *Factor* belongs to another *DNA*, *Product* or *Factor*. This relation is to be used between entities of the same nature, to denote members of a set (e.g. genes belonging to gene family, proteins to a protein family, subfamilies to families, etc.).

#### Comparison

A *DNA*, *Product* or *Factor* compared to another *DNA*, *Product* or *Factor*. This relation type is also used to link abbreviations to their full form. It has three specifications a) equivalent in function, b) identical in sequence and redundant which denotes equivalence in time, localization and function at the same time.

### Localization

#### Is Found In or During

A *Product* accumulates or is found in a given *Factor* or during a *Development Phase*. In reality this type can be directly split into two distinct types, *Is Found In* and *Is Found During*, depending on the type of the second argument, but these two types were merged in order to reduce the size of the annotation model.

### Regulation

#### Regulates Activity Of

A *DNA*, *Product* or *Factor* regulates the activity of a *Product* or a *Factor*. This relation type is used whenever it is not possible to use the more specific types *Regulates Expression Of* and *Regulates Accumulation Of*.

#### Regulates Expression Of

A *DNA*, *Product* or *Factor* regulates the activity of a *DNA* entity.

#### Regulates Accumulation Of

A *DNA*, *Product* or *Factor* regulates the accumulation of a *Product*, and more specifically of a *Protein*, *RNA*, or a *Hormone*.

#### N-ary Event

#### Condition

Condition is the only type of relation in this corpus which can take another relation as an argument. It is used to emulate multi-argument events and it is comparable to the mechanism of secondary arguments in the conceptual model.

### 2.3.3 Model Transformations

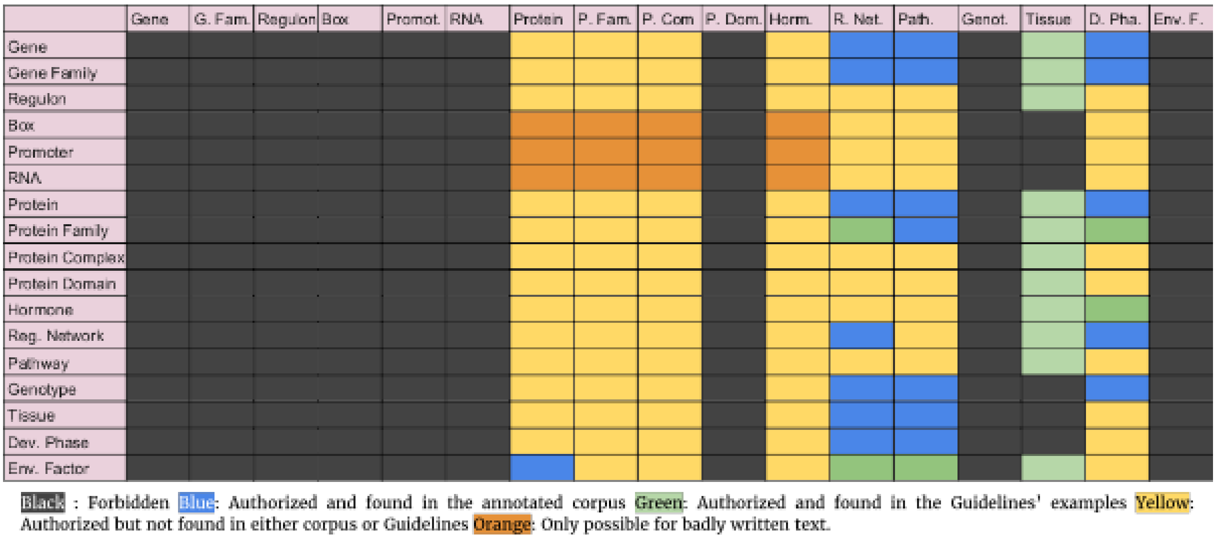
#### Annotations to Concepts

While conducting a detailed study of the possible argument combinations for the relations of the annotation model, we observed the phenomenon of clearly defined relation subtypes, like in the example of regulation illustrated in figure 2.11. This is the consequence of the decision to keep the number of relation types minimal in the annotation model, and it served as a guide for the creation of the conceptual model. Figures in 2.12 contain the three relation types of the conceptual model corresponding to Regulation of Activity in the annotation model. In this example, we follow the modification of the model thanks to the observation that there exist three semantically separable cases on the constraint table for this relation type (Fig. 2.11), guiding us to split this type into three separate relation types (Fig. 2.12).

Along with the definitions included in this thesis and the model descriptions and Guidelines documents, a detailed list of valid argument combinations, called signatures, were provided [Chaix et al., 2016c]. These argument type constraints can play an important role in relation extraction, as definitions that are too permissive can create a lot of candidate relations which lower the performance of the algorithms and increase the complexity of the learning models.

#### Transformation for Relation Extraction

The system developed for RE (AlvisRE, presented in Chapter IV) on this corpus works by extracting binary relations between *entities*. Even though the manual annotation was done



| | Gene | G. Fam. | Regulon | Box | Promot. | RNA | Protein | P. Fam. | P. Com | P. Dom. | Horm. | R. Net. | Path. | Genot. | Tissue | D. Pha. | Env. F. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟦 | 🟦 | ⬛ | 🟩 | 🟦 | ⬛ |
| Gene Family | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟦 | 🟦 | ⬛ | 🟩 | 🟦 | ⬛ |
| Regulon | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟩 | 🟨 | ⬛ |
| Box | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟧 | 🟧 | 🟧 | ⬛ | 🟧 | 🟨 | 🟨 | ⬛ | ⬛ | 🟨 | ⬛ |
| Promoter | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟧 | 🟧 | 🟧 | ⬛ | 🟧 | 🟨 | 🟨 | ⬛ | ⬛ | 🟨 | ⬛ |
| RNA | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟧 | 🟧 | 🟧 | ⬛ | 🟧 | 🟨 | 🟨 | ⬛ | ⬛ | 🟨 | ⬛ |
| Protein | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟦 | 🟦 | ⬛ | 🟩 | 🟦 | ⬛ |
| Protein Family | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟩 | 🟦 | ⬛ | 🟩 | 🟩 | ⬛ |
| Protein Complex | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟩 | 🟨 | ⬛ |
| Protein Domain | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟩 | 🟨 | ⬛ |
| Hormone | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟩 | 🟩 | ⬛ |
| Reg. Network | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟦 | 🟨 | ⬛ | 🟩 | 🟦 | ⬛ |
| Pathway | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟩 | 🟨 | ⬛ |
| Genotype | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟦 | 🟦 | ⬛ | ⬛ | 🟦 | ⬛ |
| Tissue | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟦 | 🟦 | ⬛ | ⬛ | 🟨 | ⬛ |
| Dev. Phase | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟨 | 🟨 | 🟨 | ⬛ | 🟨 | 🟦 | 🟦 | ⬛ | ⬛ | 🟨 | ⬛ |
| Env. Factor | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟦 | 🟨 | 🟨 | ⬛ | 🟨 | 🟩 | 🟩 | ⬛ | 🟩 | 🟨 | ⬛ |

Black : Forbidden Blue: Authorized and found in the annotated corpus Green: Authorized and found in the Guidelines' examples Yellow: Authorized but not found in either corpus or Guidelines Orange: Only possible for badly written text.

Figure 2.11: Constraint table for the relation type "Regulates Activity Of". Each row of the table corresponds to the type of the first relation argument, and each column to the type of the second one.

using a model based on two-argument relations, the *Condition* relation type also takes relations as arguments, in order to emulate multi-argument events. Consequently, even though it only contains binary relations, this model cannot directly be used with AlvisRE.

The problem with such s transformation is that there is no "one size fits all" solution. If we look, for example, at the relation in figure 2.13, the following problems arise:

a. (A,B) are annotated as being in a relation conditional to C, so it is not safe to automatically assume that this relation still holds without C (is this true?), or that this new relation is of the same type.

b. It is also uncertain if the pairs (A,C) and (B,C) are in a relation, and of what type.

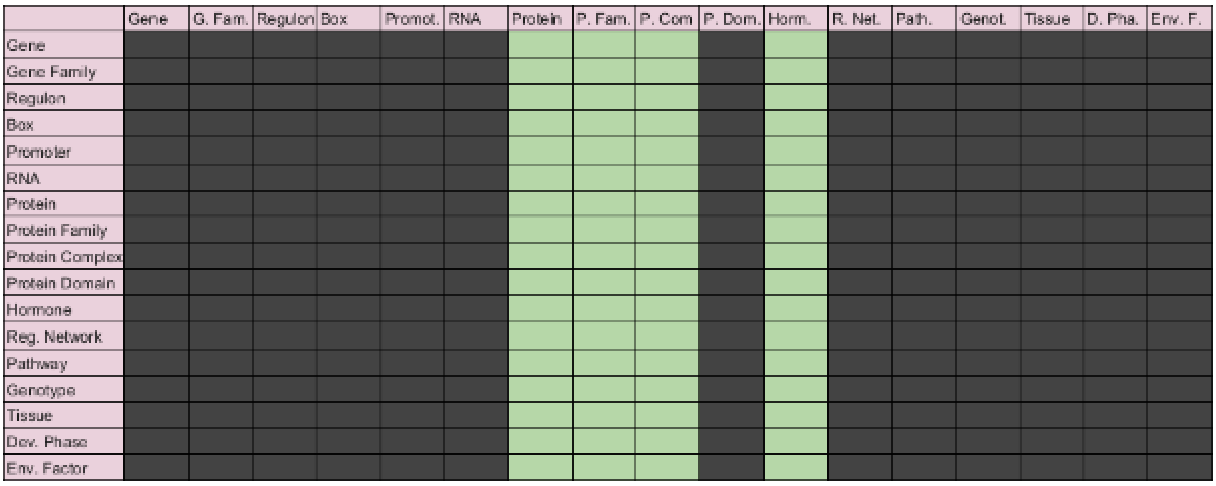
It is impossible to give a global answer to these questions that holds true for all types, so it was necessary to examine each case encountered separately and infer a set of rules. Testing all possible combinations was not a realistic goal, so the inferred rules cover the scenarios which exist in the corpus.

### Transformation Rules

While an extensive list of the rules can be found in Appendix H, some characteristic rules of transformation are listed below. In these examples, **Agent** corresponds to any of the primary arguments (A and B in Figure 2.13) and **Constraint** to the **Constraint** argument of the *Condition* relation (C in the figure). Once again, argument names are found in bold letters, whereas defined entity and relation types are in italic.

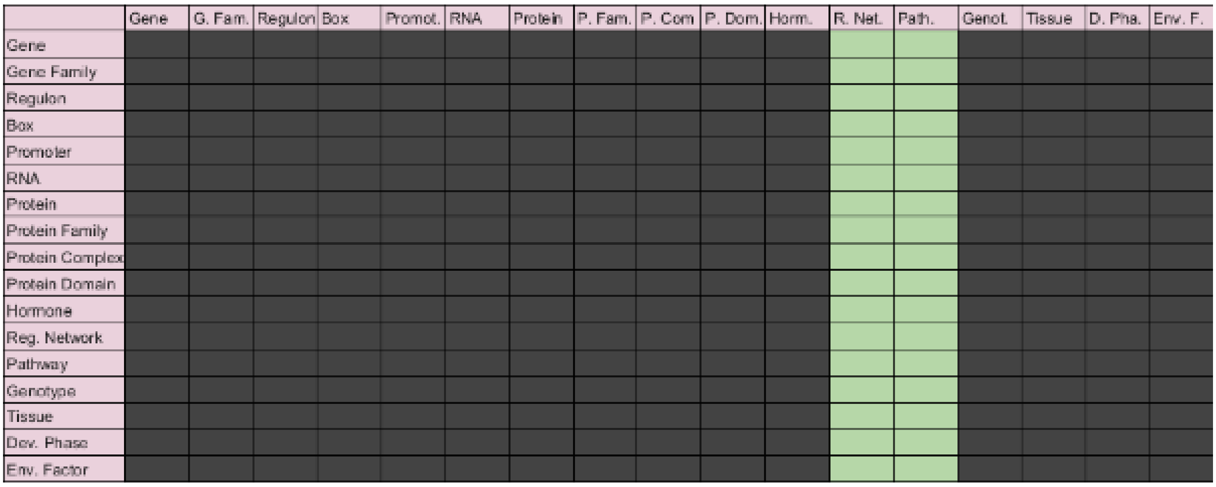
### Case: Agent ∈ { Gene, Gene Family, Box, Promoter}

In case of the relation belonging to type *Belongs To* or *Comparison*, conditional to *Genotype*, the transformation is as follows:



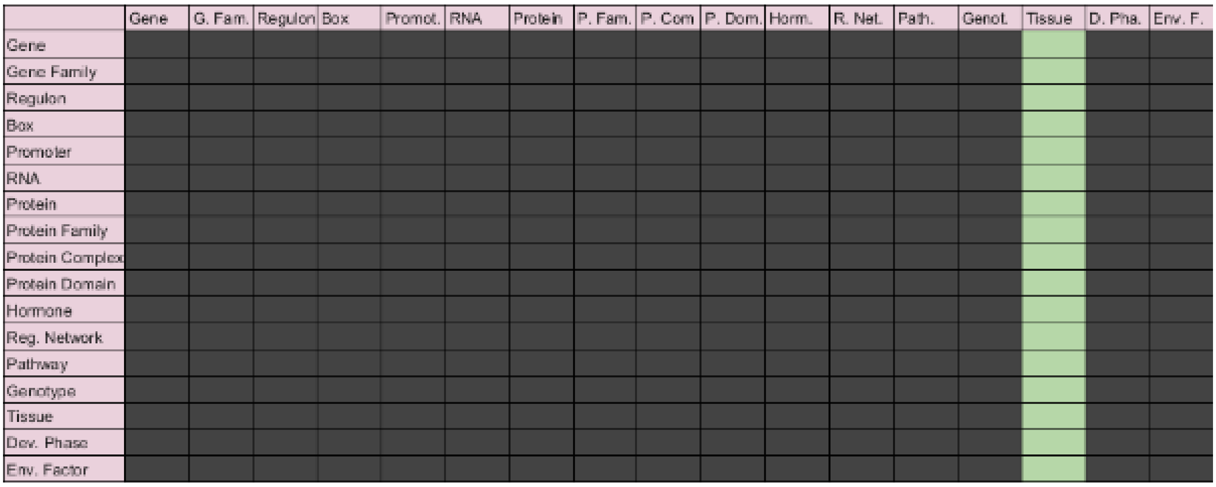
| | Gene | G. Fam. | Regulon | Box | Promot. | RNA | Protein | P. Fam. | P. Com | P. Dom. | Horm. | R. Net. | Path. | Genot. | Tissue | D. Pha. | Env. F. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Gene Family | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Regulon | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Box | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Promoter | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| RNA | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Protein | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Protein Family | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Protein Complex | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Protein Domain | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Hormone | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Reg. Network | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Pathway | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Genotype | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Tissue | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Dev. Phase | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |
| Env. Factor | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | 🟩 | ⬛ | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ |

(a) Regulation of Molecule Activity.



| | Gene | G. Fam. | Regulon | Box | Promot. | RNA | Protein | P. Fam. | P. Com | P. Dom. | Horm. | R. Net. | Path. | Genot. | Tissue | D. Pha. | Env. F. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Gene Family | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Regulon | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Box | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Promoter | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| RNA | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Protein | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Protein Family | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Protein Complex | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Protein Domain | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Hormone | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Reg. Network | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Pathway | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Genotype | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Tissue | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Dev. Phase | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |
| Env. Factor | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | 🟩 | ⬛ | ⬛ | ⬛ | ⬛ |

(b) Regulation of Process.



| | Gene | G. Fam. | Regulon | Box | Promot. | RNA | Protein | P. Fam. | P. Com | P. Dom. | Horm. | R. Net. | Path. | Genot. | Tissue | D. Pha. | Env. F. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Gene Family | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Regulon | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Box | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Promoter | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| RNA | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Protein | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Protein Family | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Protein Complex | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Protein Domain | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Hormone | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Reg. Network | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Pathway | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Genotype | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Tissue | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Dev. Phase | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |
| Env. Factor | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | ⬛ | 🟩 | ⬛ | ⬛ |

(c) Regulation of Tissue Development.

Figure 2.12: Splitting "Regulates Activity Of" into three semantically seperable relation types.

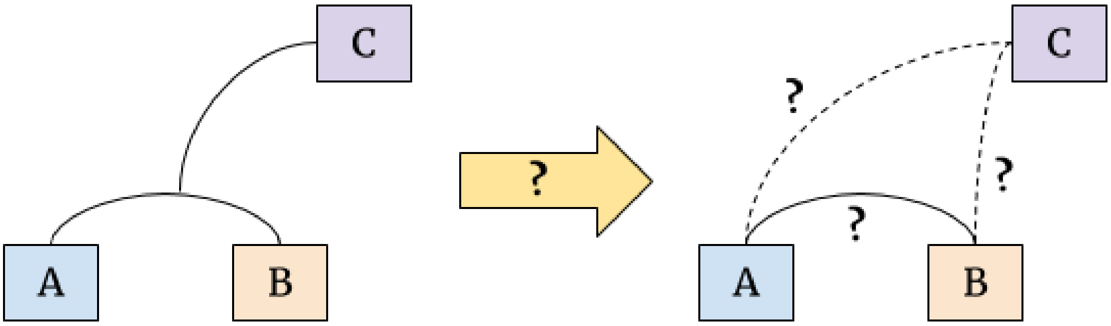


Figure 2.13: Transformation of a condition relation.

- *Gene* Belongs To *Gene Family* conditional to *Genotype* →
  - *Gene Belongs To Gene Family*
  - *Gene Is Found In Or During Genotype*
  - *Gene Family Is Found In Or During Genotype*

Otherwise:

- If **Constraint**∈ {*Genotype* , *Development Phase*, *Tissue*} →
  - **Constraint** *Regulates Expression Of* **Agent**
- If **Constraint**∈ {*Environmental Factor*} →
  - No relation with **Constraint** entity.

**Case: Agent ∈ {Protein, Protein Family, Protein Domain, Protein Complex, RNA}**

- **Constraint**∈ {*Genotype* , *Development Phase*, *Tissue*} →
  - **Agent** *Is Found In Or During* **Constraint**
- **Constraint** ∈ {*Environmental Factor*} →
  - **Constraint** *Regulates Activity Of* **Agent**

**Case: Agent∈ {Regulatory Network}**

- Regardless of type of **Constraint** →
  - **Constraint** *Regulates Activity Of* **Agent**

**Case: Agent**$\in$ **{Pathway, Tissue}**

- **Constraint**$\in$ {*Genotype*, *Tissue*} →
  - **Agent** *Is Found In Or During* **Constraint**
- **Constraint**$\in$ {*Environmental Factor*} →
  - **Constraint** *Regulates Activity Of* **Agent**
- **Agent**$\in$ {*Pathway*} and **Constraint** $\in$ {*Development Phase*} →
  - **Agent** *Is Found In Or During* **Constraint**
- **Agent**$\in$ {*Tissue*} and **Constraint** $\in$ {*Development Phase*} →
  - **Constraint** *Regulates Activity Of* **Agent**

**Case: Agent** $\in$ **{Development Phase** }

- **Constraint**$\in$ {*Genotype*, *Environmental Factor*} →
  - **Constraint** *Regulates Activity Of* **Agent**
- **Constraint**$\in$ {*Development Phase*} →
  - **Agent** *Is Found In During* **Constraint**
- **Constraint**$\in$ {*Tissue*} →
  - **Agent** *Regulates Activity Of* **Constraint**

**Case: Agent** $\in$ **{Genotype}**

- **Constraint**$\in$ {*Tissue*} →
  - **Constraint** *Is Found In Or During* **Agent**
- **Constraint**$\in$ {*Development Phase*} →
  - **Agent** Regulates Activity Of **Constraint**
- **Constraint** $\in$ {*Genotype*} →
  - **Agent** *Is Found In Or During* **Constraint**
- **Constraint** $\in$ {*Environmental Factor*} →
  - **Constraint** *Regulates Activity Of* **Agent**

## 2.4 Corpus

### 2.4.1 Source

The corpus consists of 45 scientific publications provided by the collaborating experts in an effort to reflect the literature they would normally consult on the subject. Hence, the selected articles

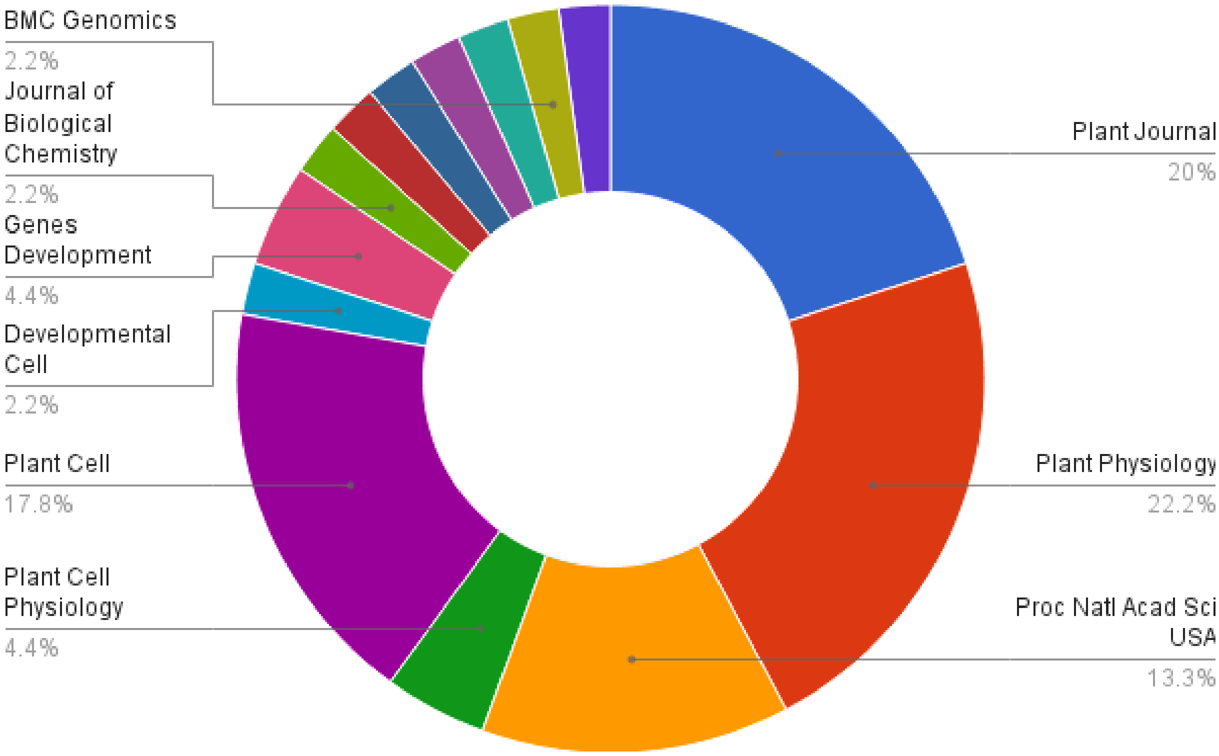


Figure 2.14: The distribution of source publications for the articles of the corpus. 13.3% only come from a conference.

come from highly respected publications in the domain and are well-cited. These articles were published between 1998 and 2012 in scientific journals (87%) and conferences (13%) (see Figures 2.14, 2.15). The vast majority (43 out of 45) of these articles are Open-Access publications, and the corpus has an average citation count of 166 (median 133).

As most of them are journal publications, their average length is that of 11 pages (Figure 2.16). However, the experts did not annotate complete documents, but selected paragraphs, based on a criterion of pertinence to the model and objectives of the task. For the 20 manually annotated documents that served for the SeeDev Challenge, the experts chose and completely annotated 87 passages of an average length of 421 words (Figure 2.18).

As seen in Figure 2.17, almost half (48%) of these selected passages came from the "Discussion" sections of the articles, while 24, 17 and 10% came from the "Introduction", "Abstract" and "Results" sections respectively. The "Discussion" sections in biology publications summarize, contextualize and interpret the findings in detail, thus making them the ideal candidates for IE tasks. Discussion was consistently annotated in all of the 20 articles. Similarly, introductions and abstracts also offer a summarization but often lack in detail and context, making them interesting but often less rich in information. Results, finally, while containing a lot of detail, often lack context and so even though they make good candidates for the annotation of named entities, they don't always provide enough information on relations.

This set was split in 6 annotation packs randomly. Packs 1-3 were selected to be used for the BioNLP '13 challenge and were adjudicated and curated accordingly. My experiments (chapter VI) are based on these three packs. The complete document list can be found in Appendix D.

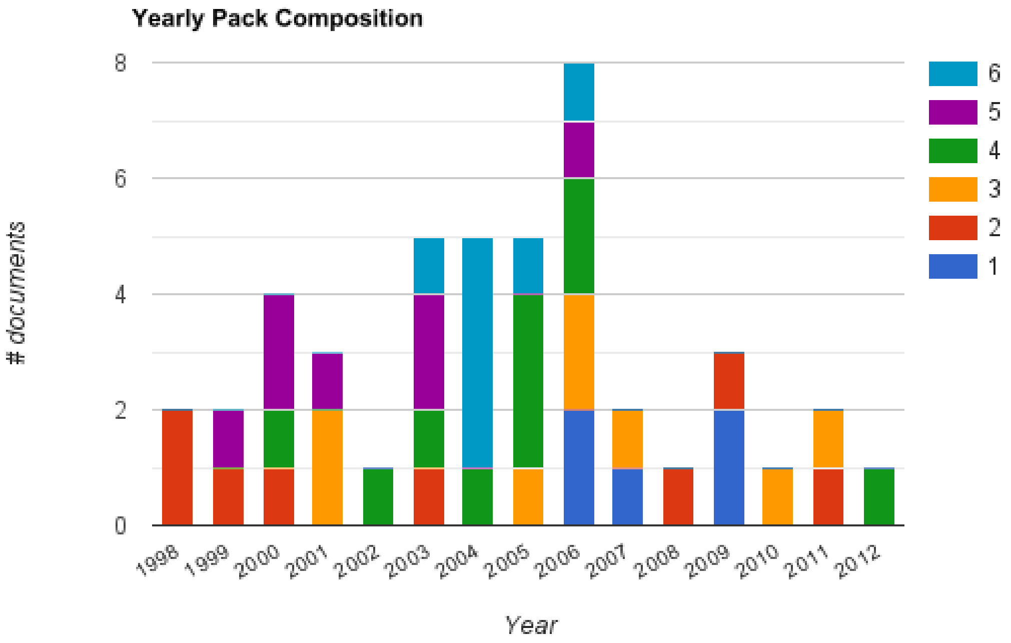


Figure 2.15: Yearly distribution of the document packs.

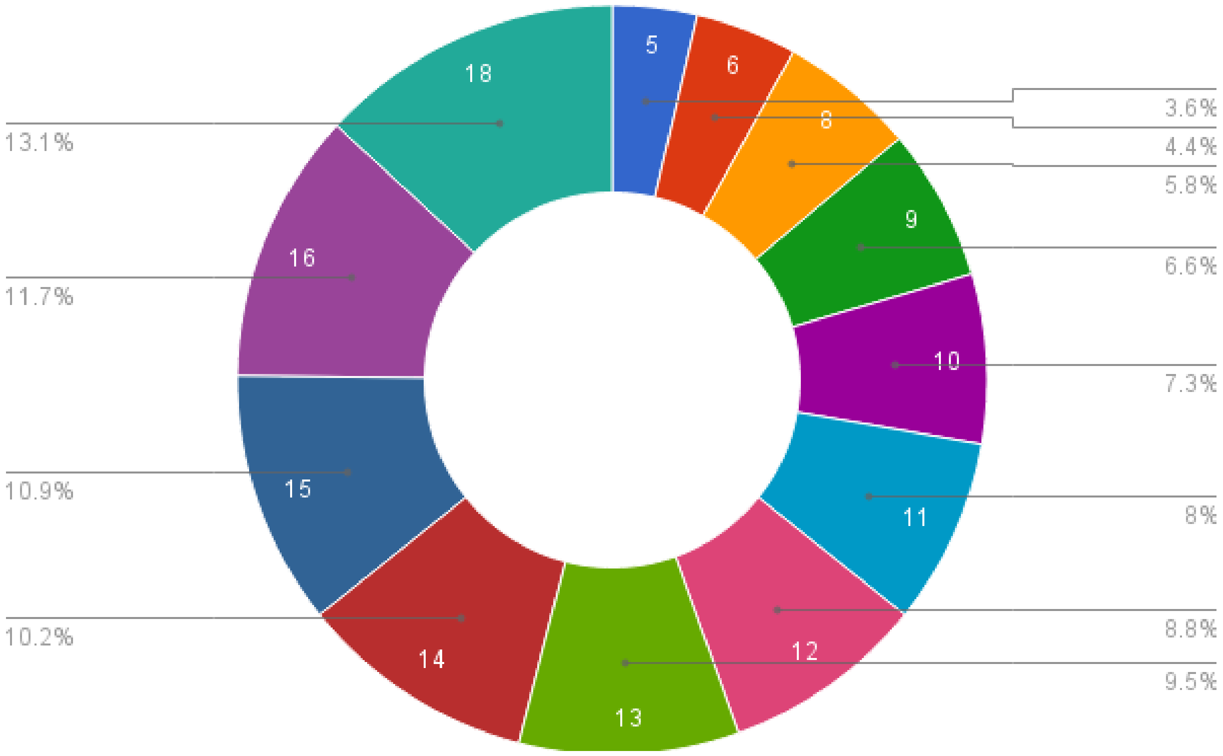


Figure 2.16: Document length (in pages) in the corpus. Longer journal articles are more frequent.

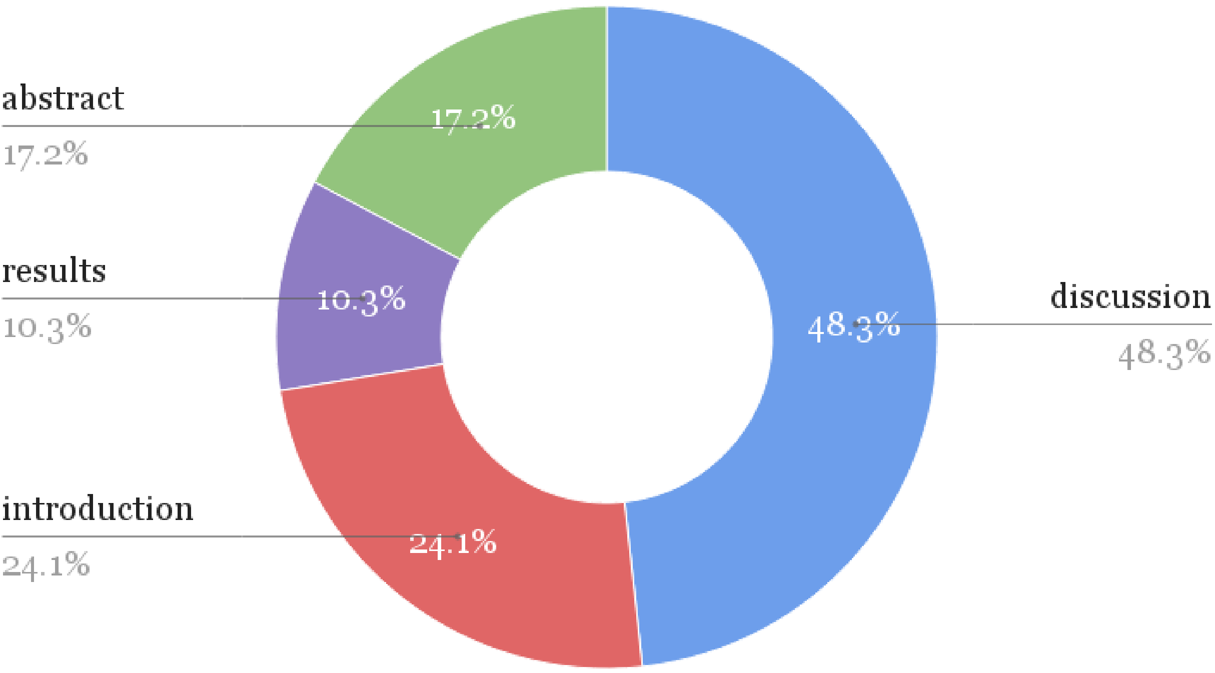


Figure 2.17: The distribution of the sections to which the select passages belong.

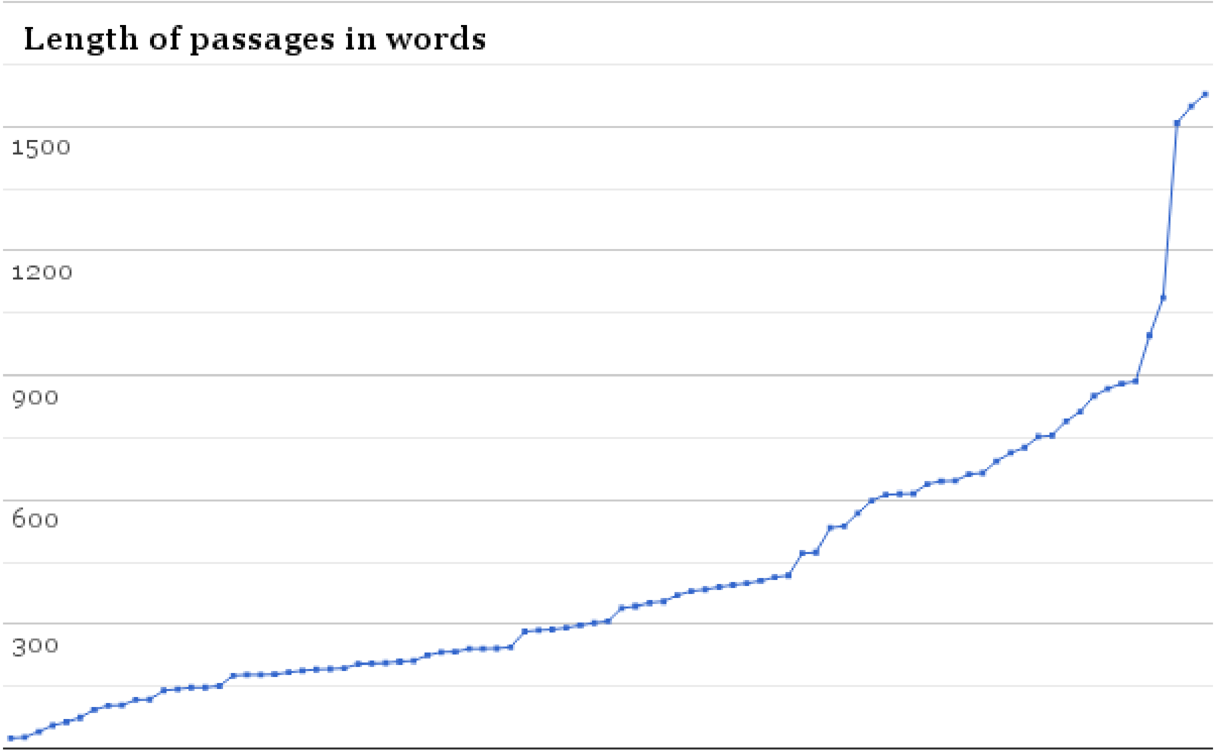


Figure 2.18: The length of the selected fully annotated passages, with an average of 421 and a median of 307 words.

The six annotation packs are as follows:

- Pack 1 : documents 4, 8-10, 12 in the order listed in App. D.
- Pack 2: documents 2, 29, 32, 33, 36, 39, 42, 44
- Pack 3: documents 28, 30, 34, 35, 40, 41, 43, 45
- Pack 4: documents 3, 5, 6, 11, 13, 16, 22, 25, 26, 31
- Pack 5: documents 1, 14, 18, 21, 23, 24, 27
- Pack 6: documents 7, 15, 17, 19, 20, 37, 38

### 2.4.2 Corpus Annotation

#### Automatic pre-annotation

Pre-annotation has been shown to reduce the time consumed for manual annotation without impacting accuracy [Lingren et al., 2014]. The experience of the annotating experts was positive, as they found that this preprocessing substantially sped up their work. Even in the cases where there were mistakes, pre-annotations draw the annotator's attention and it is improbable that they go unnoticed.

AlvisNLP/ML was used for named entity recognition, and more specifically, a modified dictionary projection method. This modified NER method takes into account the typography of the text, as this helps disambiguate between certain entity types in the case of *A. Thaliana*. The dictionaries for each type were the following:

- Box: Agris[11] [Yilmaz et al., 2011, Davuluri et al., 2003] and NIAS DNABank[12]
- Genes, Proteins and families: TAIR [13]
- Pathway: Plant_ Pathway and Virus_Pathway from TAIR
- Tissue : manually[14] constructed dictionary of 25 terms
- Development Phase : manually[1] constructed dictionary of 16 terms

#### Campaigns

The annotation was organized in packs and campaigns. At the time of writing this thesis 3 out of the initial 6 packs have been completed: packs 1, 2 and 3. The first pack was the smallest one and it was the one used to develop and validate the Guidelines document and the annotation model, it was annotated mainly by one annotator. The other two packs were annotated in a more elaborate, double blind fashion and adjudicated. In summary:

[11] http://arabidopsis.med.ohio-state.edu/

[12] http://www.dna.affrc.go.jp/

[13] http://arabidopsis.org

[14] By the team members with biology backgrounds.

| Annotator | PRE | REC | F-M |
|---|---|---|---|
| A | 79.8 | 41.7 | 54.8 |
| B | 75.4 | 57.5 | 65.3 |
| $A \cup B$ | 72.8 | 72.0 | 72.4 |

Table 2.1: Inter-annotator agreement, evaluated by comparing each annotator output to the reference annotation.

### Pack 1

- Entities pre-annotated by AlvisNLP/ML
- Entities and relations annotated by Bertrand
- Corrections by Claire, Louise and Estelle
- Quality checked by Estelle

### Packs 2 & 3

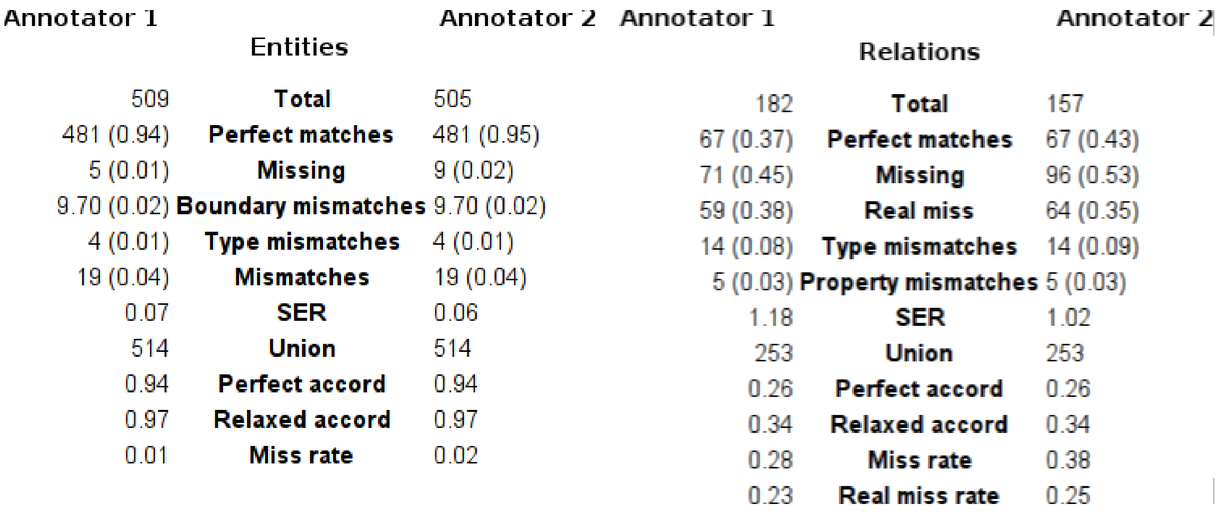
- Entities pre-annotated by AlvisNLP/ML
- Entities manually annotated by Estelle
- Double blind annotation of relations by Bertrand and Abdou
- Pre-adjudication by Estelle
- Adjudication by Estelle, Bertrand and Abdou
- Quality checked by Estelle

### Adjudication

Before adjudication one document has between 100 and 600 points to verify, including entities, relations, conditions. After a semi-automatic process of validating entities and relations which are strictly identical, there are 60-300 points remaining. This process took about half a day per document for one person (Estelle). Finally, after this stage the final adjudication took about half a day in a collaboration between 3 people (Estelle, Abdou, Bertrand). Figure 2.19 contains an example of the agreement metrics calculated by the diff tool used in the adjudication process. The result of the adjudication process was the consensus annotation set, which is used as the gold standard for this corpus and the derived corpora.

The evaluation of inter-annotator agreement was done by comparing the annotations of each annotator and the gold standard annotations and calculating recall, precision and F-measure (see Table 2.1). The differences between the individual annotators vary according to the event types[15].

[15] The recent SeeDev challenge overview publication for the BioNLP '16 Workshop [Chaix et al., 2016a] contains a more detailed analysis of inter-annotator agreement on this corpus.



| Annotator 1 | Entities | Annotator 2 |
|---|---|---|
| 509 | **Total** | 505 |
| 481 (0.94) | **Perfect matches** | 481 (0.95) |
| 5 (0.01) | **Missing** | 9 (0.02) |
| 9.70 (0.02) | **Boundary mismatches** | 9.70 (0.02) |
| 4 (0.01) | **Type mismatches** | 4 (0.01) |
| 19 (0.04) | **Mismatches** | 19 (0.04) |
| 0.07 | **SER** | 0.06 |
| 514 | **Union** | 514 |
| 0.94 | **Perfect accord** | 0.94 |
| 0.97 | **Relaxed accord** | 0.97 |
| 0.01 | **Miss rate** | 0.02 |

| Annotator 1 | Relations | Annotator 2 |
|---|---|---|
| 182 | **Total** | 157 |
| 67 (0.37) | **Perfect matches** | 67 (0.43) |
| 71 (0.45) | **Missing** | 96 (0.53) |
| 59 (0.38) | **Real miss** | 64 (0.35) |
| 14 (0.08) | **Type mismatches** | 14 (0.09) |
| 5 (0.03) | **Property mismatches** | 5 (0.03) |
| 1.18 | **SER** | 1.02 |
| 253 | **Union** | 253 |
| 0.26 | **Perfect accord** | 0.26 |
| 0.34 | **Relaxed accord** | 0.34 |
| 0.28 | **Miss rate** | 0.38 |
| 0.23 | **Real miss rate** | 0.25 |

Figure 2.19: Annotator agreement as calculated by the diff tool.
**Slot Error Rate (SER):** 0 is perfect and 1 is total mismatch.
**Union:** Perfect Matches + Miss Annotator 1 + Miss Annotator 2 + Type Mismatches + Property Mismatches (for relations).
**Perfect Accord:** Ratio Perfect Matches / Union
**Relaxed Accord:** Ratio (Perfect Matches + Type + Property) / Union
**Miss Rate:** Miss Annotator 1 or 2 / Union
**Real Miss Rate:** Real Miss (see above) of Annotator 1 or 2 / Union

## 2.5 Results and Discussion

The resulting corpus consists of twenty documents, split into 87 completely annotated passages. The corpus contains 7082 entity annotations and 3923 relation annotations. Figures 2.20 and 2.21 represent the distribution of entity and relation annotations in the corpus respectively.

As far as entities are concerned, the most predominant types are Tissue (18%), Protein (16%), Gene (12%), Genotype (12%) and Development Phase (11%). Genes, genotypes and proteins are fundamental in regulatory networks, which explains their frequency in articles of this domain. Additionally, they are present as required arguments in most relation types, making them indispensable. Tissue and Development Phase are the most common conditions for any relations, and required arguments for Localization, with Condition and Localization (Is Found In or During) being two of the most commonly occurring types in the corpus (28% and 18% respectively), along with the Regulation family (33%).

Lesser common types of entities fall into three general categories: a) more technical types (Box, Domain, etc) expected to be found mostly in “Results” sections of articles, which were less often selected by the experts for annotation, b) more complex types, like Regulatory Network or Pathway, which in spite of their importance in this topic’s literature, are not always explicitly mentioned in the text, and c) other types, like Environmental Factor and Hormone, which simply are less commonly present in these texts.

As mentioned earlier, Localization and Regulation make up for the majority of relation annotations (60% in total), which is to be expected, as these types are quintessential to regulatory networks.

| | # | Train | Dev | Test |
|---|---|---|---|---|
| Articles | 20 | 90% | 75% | 80% |
| Passages | 87 | 45% | 22% | 33% |
| Words | 44857 | 45% | 23% | 33% |
| Total entity annotations | 7082 | 46% | 23% | 31% |
| Total n-ary relation annotations (SeeDev full) | 2583 | 45% | 23% | 32% |
| Total binary relation annotations (SeeDev binary) | 3575 | 46% | 23% | 32% |

Table 2.2: A summarized table of the distribution of the SeeDev corpus in the train, dev, and test sets. As the articles were split into passages the first row corresponds to the coverage of source articles in the corresponding sets, contrary to the rest of the table which should be read as percentage split among the sets.

Conditions account for 28% of annotations, something that has more to do with the choices made in the annotation model than the domain itself: these conditions are semantically part of other relations as they provide a mechanism to add additional arguments while keeping the annotation model two-argument only, for practical reasons.

Additionally to the original annotated corpus which corresponds to the annotation model presented in this chapter, this corpus is available in the following forms, transformed according to the processes described in section 2.3.3.

1. The binary version of the annotated corpus, according to the annotation model presented here, but with a transformation of the condition relations, so as to only contain binary entity-entity relations.
2. The SeeDev full corpus, which corresponds to the conceptual model and was used for the BioNLP 2016 Shared Task[16].
3. The Seedev binary corpus, which also corresponds to the conceptual model, but is transformed so as to only contain binary entity-entity relations. This version was equally proposed in the BioNLP 2016 SeeDev challenge.

For the SeeDev corpus, as it was used for the BioNLP challenge, some analytical results are presented in this work. The corpus was split in three sets, training, development and test for use by the participants of the challenge. The first two sets were available labeled for training purposes, and the third one was only published unlabeled, for evaluation. Table 2.2 shows the counts, as well as the percentages of source material and annotations in these three sets.

While the entity types are the same between the conceptual and the manual annotation model, relation type statistics changed from those presented above in Figure 2.21. Table 2.3 shows the occurrences of relation annotations in the SeeDev corpus across the three datasets, and Figure 2.22 illustrates the overall distribution of the relation annotations, in a manner analogous to Fig. 2.21. In this version, too, the Regulation family is the most prevalent one, totaling approx. 45% of annotations. Their actual number is unchanged, but the change in percentage is due to the fact that the Condition type disappears in this model. The relation type family Time and Localization was most often used to represent conditions, and this is why it is the second most frequent one, just as Localization was in the manually annotated corpus. Most of the relation types are proportionally found in the train, dev and test sets, with the exception of the following:

[16] http://2016.bionlp-st.org/tasks/seedev

| Relation Type Family | # | Train | Dev | Test | Total |
|---|---:|---:|---:|---:|---:|
| **Time and Localization** | **704** | **45%** | **23%** | **32%** | **20%** |
| Exists At Stage | 33 | 45% | 24% | 30% | 1% |
| Exists In Genotype | 377 | 45% | 21% | 34% | 11% |
| Occurs During | 30 | 27% | 33% | 40% | 1% |
| Occurs In Genotype | 48 | 38% | 33% | 29% | 1% |
| Is Localized In | 216 | 50% | 22% | 29% | 6% |
| **Function** | 257 | **42%** | **28%** | **30%** | **7%** |
| Is Involved In Process | 55 | 42% | 36% | 22% | 2% |
| Transcribes Or Translates To | 54 | 46% | 24% | 30% | 2% |
| Is Functionally Equivalent To | 148 | 41% | 26% | 33% | 4% |
| **Regulation** | **1731** | **46%** | **22%** | **31%** | **48%** |
| Regulates Accumulation | 81 | 44% | 36% | 20% | 2% |
| Regulates Development Phase | 242 | 44% | 24% | 32% | 7% |
| Regulates Expression | 450 | 45% | 25% | 31% | 13% |
| Regulates Molecule Activity | 25 | 64% | 0% | 36% | 1% |
| Regulates Process | 904 | 48% | 20% | 32% | 25% |
| Regulates Tissue Development | 29 | 31% | 31% | 38% | 1% |
| **Composition and Membership** | **532** | **44%** | **22%** | **34%** | **15%** |
| Composes Primary Structure | 51 | 39% | 29% | 31% | 1% |
| Composes Protein Complex | 19 | 84% | 0% | 16% | 1% |
| Has Sequence Identical To | 126 | 49% | 16% | 35% | 4% |
| Is Member Of Family | 230 | 39% | 24% | 37% | 6% |
| Is Protein Domain Of | 106 | 43% | 27% | 29% | 3% |
| Interaction | 264 | 46% | 21% | 33% | 7% |
| Interacts With | 148 | 42% | 22% | 36% | 4% |
| Binds To | 116 | 52% | 21% | 28% | 3% |
| **Specific to Binary Framework** | **87** | **51%** | **26%** | **23%** | **2%** |
| Is Linked To | 87 | 51% | 26% | 23% | 2% |
| **Total** | **3575** | **46%** | **23%** | **32%** | **100%** |

Table 2.3: The distribution of relation types in the SeeDev Challenge corpus. Train, Dev and Test column percentages represent the distribution of each type in the three datasets. The column Total represents the distribution of each type in the whole corpus.

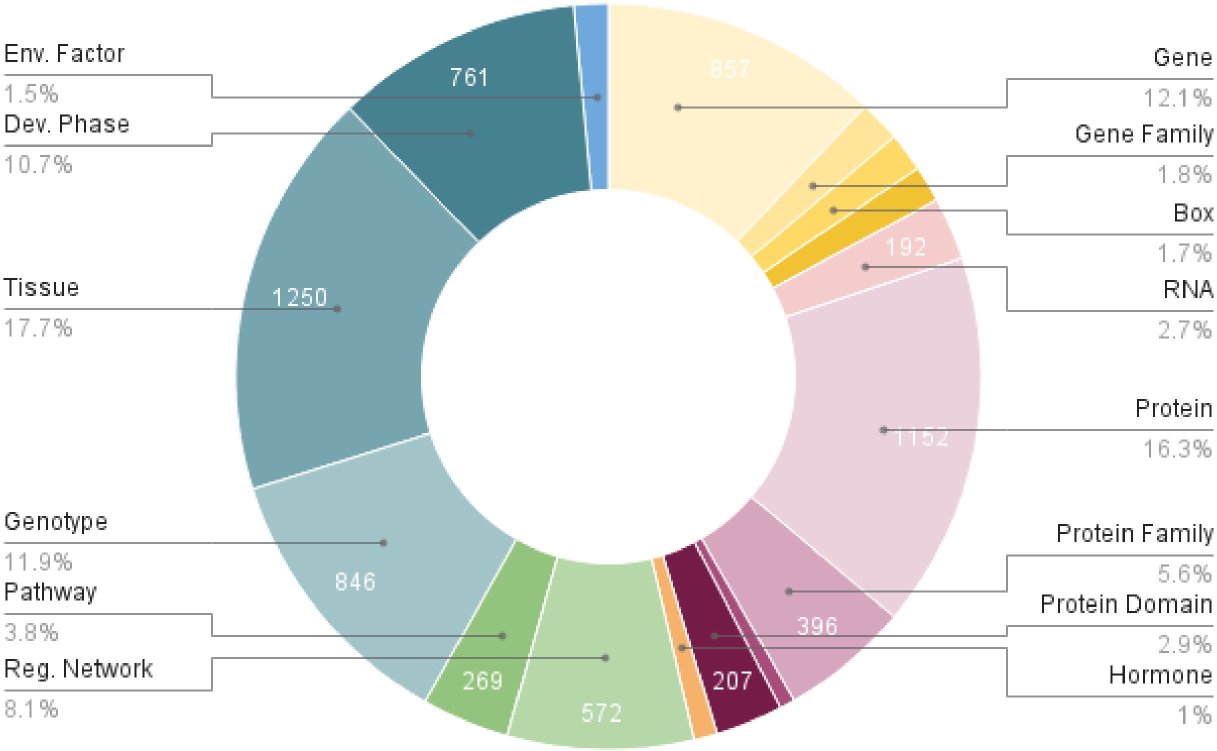


Figure 2.20: Distribution of entity annotations in the manually annotated corpus (packs 1-3).

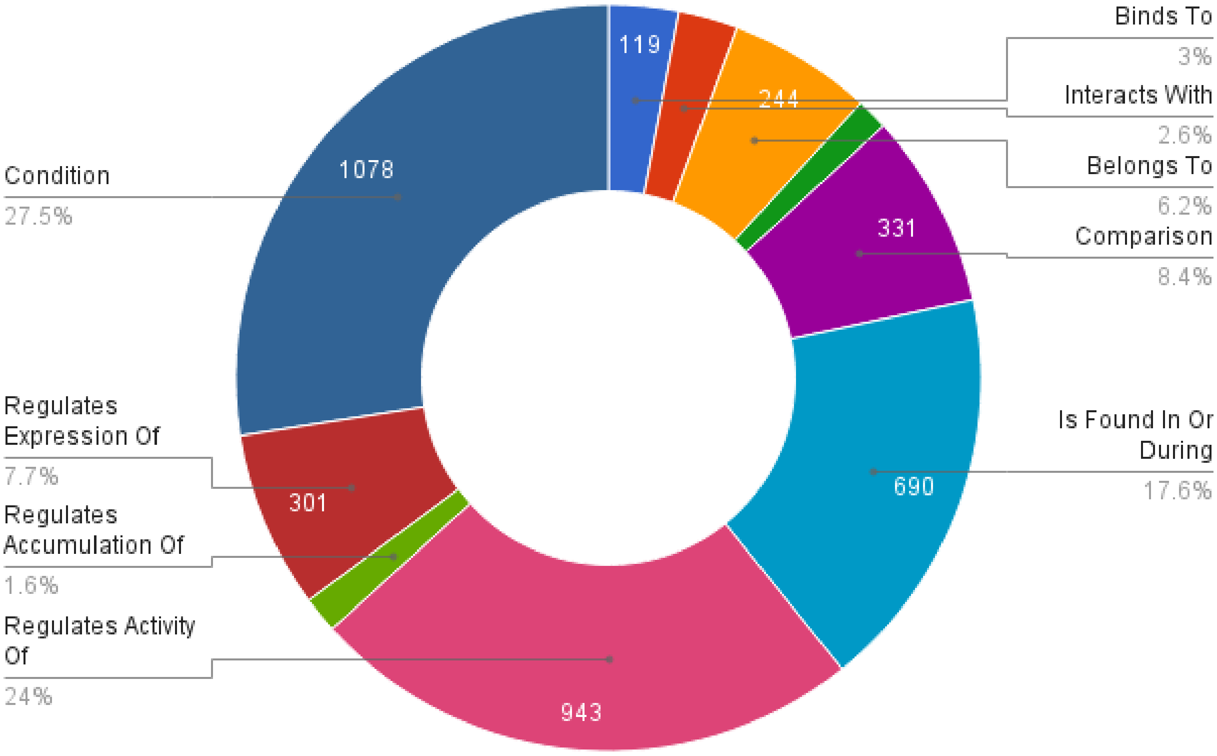


Figure 2.21: Distribution of relation annotations in the manually annotated corpus (packs 1-3).

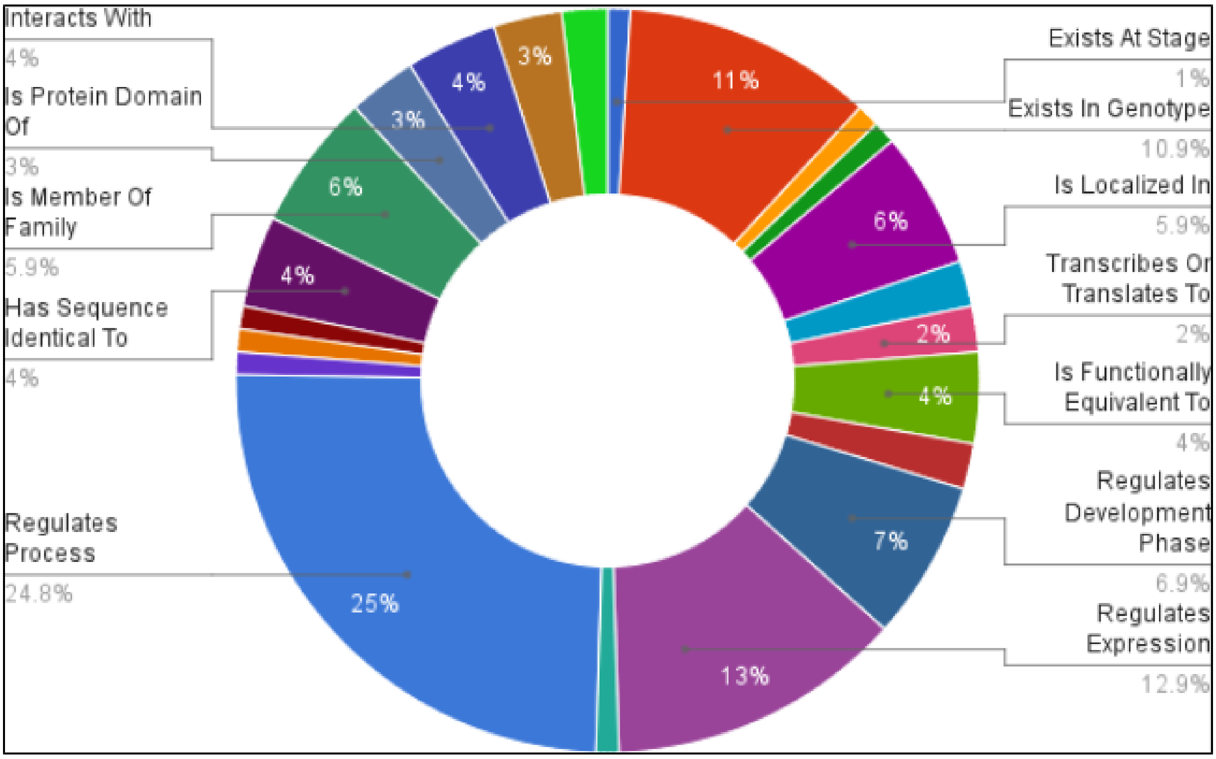


Figure 2.22: Distribution of relation types in the transformed SeeDev corpus.

1. 40% of all Occurs During relations are found in test. Combined with the fact that only 30 Occurs During relations occur in the entire corpus, this could lead to prediction difficulties.
2. No occurrences of Regulates Molecule Activity and Composes Primary Structure are found in the dev set, which serves for testing during development. While this could be problematic for model building, cross validation would solve this issue. Since this phenomenon occurs in the dev set, the participants are aware of it during development and can address it.

A number of relation types are scarce in the corpus (see: <5% in the last column of Table 2.3). Depending on the heterogeneity of expression for these types in the corpus it could be difficult to predict them with such a low number of examples.

## 2.6 Conclusion

By producing two models, the dual goal of producing a complete and precise knowledge model, and a model that is adapted to machine learning predictions was accomplished. The models and corpora produced for A. Thaliana are the product of years of collaboration from a team of experts, resulting in high sophistication, quality and pertinence for the biology domain. Their documentation is rich, they are the subject of a number of already published and upcoming scientific works and they are available online. Both models and corpora were produced with an emphasis on quality and reusability. These facts make them ideal for reuse by both the IE and biology community, as they can serve as good benchmarks for the first and a sound basis for applications and collaborations for the second.

My contributions to the work presented in this chapter were mainly in model design and documentation, but I also helped during the annotation campaigns and presented our work on various occasions.

# Chapter 3

# Relation Extraction

**3.1 Introduction . . . 78**
3.1.1 Information Extraction . . . 79
Named Entity Recognition . . . 79
Coreference Resolution . . . 79
Relation Extraction . . . 80
**3.2 Linguistic Pre-processing . . . 81**
3.2.1 Tokenization and segmentation . . . 81
3.2.2 Lemmatization and normalization . . . 81
3.2.3 Syntactic Parsing . . . 81
**3.3 Representation . . . 82**
3.3.1 Introduction . . . 82
3.3.2 From text files to complex sentence objects . . . 82
The file format . . . 83
AlvisRE object architecture . . . 84
3.3.3 From sentences to candidate relations . . . 84
Calculating candidates by co-occurrence . . . 85
3.3.4 From candidate relations to paths . . . 85
Introduction . . . 85
Dependency Paths . . . 86
Word-Order Paths . . . 89
Combined Paths . . . 89
3.3.5 From paths to a machine-learning ready representation . . . 91
Introduction . . . 91
Co-occurrence . . . 91
Paths as sequences . . . 91
Global Alignment and a similarity-based representation . . . 92
Example . . . 95

Vector transformation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95
Using varied sources of information . . . . . . . . . . . . . . . . . . . . . . 96
Algorithmic complexity . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
**3.4 Classification . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98**
3.4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
3.4.2 Support Vector Machines . . . . . . . . . . . . . . . . . . . . . . . . 99
**3.5 Adding Semantic Information . . . . . . . . . . . . . . . . . . . . . . . 101**
3.5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 101
3.5.2 Manual Classes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
3.5.3 WordNet . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
3.5.4 Textual similarity . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103
3.5.5 Distributional Semantics . . . . . . . . . . . . . . . . . . . . . . . . . 103
Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103
DISCO . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104
Word Embeddings . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105
Combining Sources . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105
**3.6 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106**
3.6.1 Experimental setup . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
Alvis Suite . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
3.6.2 Experimental validation . . . . . . . . . . . . . . . . . . . . . . . . . 106
A fix for missing dependency links. . . . . . . . . . . . . . . . . . . . . 108
Representation and Similarity Function . . . . . . . . . . . . . . . . . . 110
Semantic Information . . . . . . . . . . . . . . . . . . . . . . . . . . 112
A complete IE task . . . . . . . . . . . . . . . . . . . . . . . . . . . . 114
**3.7 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116**
**3.8 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 117**

## 3.1 Introduction

Information Extraction (IE) aims to extract structured information from unstructured text in natural language. This thesis focuses on Relation Extraction (RE), but it is part of a larger work on biomedical IE. In the spirit of giving a complete picture, the other components and choices will be described briefly.

The relation extraction system proposed here is named AlvisRE and is part of the Alvis Suite [Ba and Bossy, 2016, Nédellec et al., 2009]. In the information extraction system described here, Named Entity Recognition (NER) and coreference resolution (CR) are done either automatically by AlvisNLP/ML components, or manually by domain experts, but always given as an input. AlvisRE is tasked with Relation Extraction (RE). It follows a similarity-based machine learning approach.

This chapter will focus on Relation Extraction and AlvisRE. It will detail its input, transformations, representation and prediction methods. Different problems, ideas and hypotheses will be presented and explored thematically. The last section of this chapter is dedicated to experimental validation of these theses, and selected results will be presented in order to guide the reader through the development of AlvisRE.

Named Entity Recognition and Coreference Resolution will be briefly presented in this section, and additional details will be given whenever necessary in the following sections in order to fully describe the underlying assumptions and facts relevant to AlvisRE and the task of RE.

### 3.1.1 Information Extraction

The different subtasks of IE are described in the order they are performed, starting from NER, then Coreference Resolution, and Relation Extraction. Text pre-processing is presented in section 3.2.

#### Named Entity Recognition

Named Entity Recognition is the task of identifying in the text the Named Entities defined by the annotation schema. This includes detecting the existence of a NE and finding correct textual boundaries, as well as its classification as the correct NE type. In international challenges often IE tasks are split into separate subtasks, with RE being performed on given NEs. In lack of manually pre-annotated NEs, both tasks are necessary and can be performed in parallel or sequentially.

For the different types of corpora on which AlvisRE was used, the source of annotated entities has varied. In the LLL corpus [Nédellec, 2005b], entities manually annotated by experts are available. The second task of BioNLP '13 on Bacteria Biotopes (BB) focused on RE and provided manually annotated NEs [Bossy et al., 2015].

In the cases where AlvisRE was used in combination with NER systems, the AlvisNLP/ML pipeline was used. This was namely the case for the BioNLP '13 BB Task 3. For the prediction of named entities, AlvisNLP/ML includes methods based on dictionary projections, term analysis and machine learning [Nédellec et al., 2009, Ba and Bossy, 2016].

In the case of the BB Corpus (BioNLP '13 BB Task 3), AlvisRE used entities predicted by an elaborate NER system. Biotope habitat recognition was done using the ToMap Method [Ratkovic et al., 2012, Golik et al., 2012a], in combination with BioYaTeA [Golik et al., 2013]. For geographical entities the Stanford NER tool [Finkel et al., 2005b] was used. Finally, for the detection of Bacteria the NCBI list of prokaryotes was used along with some patterns that capture acronyms and adjust boundaries [Ratkovic et al., 2012]. More information on the particularities of the corpus and the tools can be found in [Ratkovic, 2014].

#### Coreference Resolution

Coreference in the context of IE describes the case where two words, phrases or, generally, parts of text refer to the same Named Entity. It specifically and uniquely targets the entity types that are included in the schema. The term coreference resolution is used when linking pronouns or nouns –the *anaphors*, and the NE that they refer to, the *referent.* Additionally, NE mentions can also be anaphors, as different occurrences in text can refer to a single entity. For example, in a text describing *Bacillus Subtilis*, there will be many mentions of the bacteria, but all of them refer to the same entity. In this case, it is customary to consider the first mention as the referent, and all subsequent mentions as anaphors.

In the majority of cases where AlvisRE has been used, no coreference resolution (CR) was performed. This was true for the LLL corpus, where the input is in the form of isolated sentences and, consequently, there are no coreferences spanning multiple sentences. Since AlvisRE works on a sentence level, it manages to extract these relations, even without intra-sentence coreference resolution. But in cases such as the Arabidopsis corpus where inter-sentence relations are rare (4% for the test set, for example), it can be a safer compromise to ignore the instances which need CR, as CR can introduce errors.

For the Bacteria Biotope (BB) corpus –where approximately 30% of the relations span multiple sentences, a rule-based solution was used [Ratkovic et al., 2012]. This system covers the case of grammatical anaphora and a custom list of words specific to BB, which were good candidates of being anaphors. In both cases, the anaphor occurs after the antecedent and, in short, the detected anaphors are linked to their closest possible antecedent. Zorana Ratkovic details the method used on the BB corpus for coreference resolution in her thesis [Ratkovic, 2014].

### Relation Extraction

The simplest solution in the simplest case is to predict a relation whenever two candidate arguments co-occur in a sentence (called learning instances in the following). AlvisRE implements the baseline scenario of co-occurrence, as well as a ML approach making use of various possible combinations of linguistic information. The ML approach is based on the notion of similarity, as it uses a function that calculates the similarity of couples of learning instances using their context as their descriptions. This similarity function is in the spirit of kernel methods, and the classifier used is a Support Vector Machine (SVM).

The similarity function used is known as Global Alignment and it is related to the string similarity function known as Edit Distance. Learning examples (or instances) are represented as sequences, or non-directed graph paths. In the dependency graph, nodes are words and the edges represent either syntactic relations or word contiguity. For each graph node, different levels of linguistic information can be taken into account.

Along with Named Entity Recognition and Coreference Resolution, AlvisRE expects certain linguistic analysis steps to have taken place. This pre-processing is detailed in section 3.2 of this chapter. A format based on the “BioNLP” format (which, in turn, is based on the “Genia” format) is used for AlvisRE’s input. An extension of the “BioNLP” format took place in the beginning of this work as it was deemed necessary in order to facilitate the transfer of a maximum of information from previous processing to AlvisRE. The building blocks of this format will be presented in section 3.3, titled “Representation”, which details all the necessary transformations of this input until it is used for classification. Section 3.4, titled “Classification”, describes the machine learning algorithm used for the prediction of relations. Section 3.5 explores different sources of semantic information that were considered and integrated into AlvisRE. Finally sections 3.6 and 3.7 are dedicated to experiments.

## 3.2 Linguistic Pre-processing

AlvisRE requires that the following steps have always taken place in a pre-processing phase. While a detailed account of this linguistic analysis and choices can be found in [Ratkovic, 2014], an outline of the necessary steps is given for context.

### 3.2.1 Tokenization and segmentation

Tokenization and segmentation play an important role in the performance of relation extraction in the biomedical domain [Jiang and Zhai, 2007], as they affect both the POS tagging and parsing steps.

Tokenization is based upon the notion of tokens. A *token* is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. In this work a token is considered not only as an atomic unit, but also as a more linguistically motivated basic unit of meaning (or word). Cases where the tokens are not words include names of proteins or species including whitespace (e.g. "Bacillus subtilis"), names of proteins and genes containing punctuation or numbers (e.g. "sigma (A)" and "A. thaliana") and other special cases such latin names, numbers, DNA sequences, etc (see [Ratkovic, 2014] for more details).

The AlvisNLP/ML pipeline contains two tokenization/segmentation modules: WoSMIG and SeSMIG, for word and sentence segmentation respectively. They rely on a combination of regular expression rules, domain dictionaries, heuristics and semantic annotation rules.

### 3.2.2 Lemmatization and normalization

Both lemmatization and (token) normalization refer to processes that group together different forms of the same linguistic object, so that they can be analyzed as identical afterwards. Lemmatization does so by grouping different inflected forms of a word and returning the base or dictionary form of a word, known as the lemma. Stemming is a simplistic approach for the reduction of derivation and lemmatization, as it consists of chopping off the ends of words heuristically.

Normalization groups tokens which match despite superficial differences in their character sequence and returns a canonical form (e.g. "Arabidopsis thaliana", "A. thaliana" and "Arabidopsis Thaliana").

While lemmatization and (surface) normalization are not obligatory, they can greatly influence the performance of AlvisRE, as it is directly dependent on the ability to correctly calculate the similarity between tokens, as will be shown later in this section.

### 3.2.3 Syntactic Parsing

Syntactic parsing is optional in some representation alternatives, as will be illustrated below, but AlvisRE was built with dependency-based parsing in mind. Zorana Ratkovic [Ratkovic, 2014] tested three parsers and chose to integrate CCG [Rimell and Clark, 2008] in AlvisNLP/ML. In that work, AvisRE has been tested with CCG (standard and transformed by Alvis Grammar) as well as Enju [Miyao and Tsujii, 2005]. These tests have shown that CCG (standard or optimized) produces the best input for AlvisRE. Constituent-based parsers and other dependency-based parsers have not been tested at this time.

## 3.3 Representation

### 3.3.1 Introduction

In this chapter, we will follow how text is represented and transformed throughout its processing by AlvisRE. This section explains these transformations by example, essentially detailing the way AlvisRE works.

Text, sentences and semantic relations are transformed and represented in various ways:

- as objects or data structures (Computer Science)
- as graphs, for example the syntactic trees (NLP)
- as sets of words, for example the bag-of-words representation (NLP)
- as sets of features, or attributes (Machine Learning)

All of the above representations are used by AlvisRE in different steps. After the previous steps (NER, CR, pre-processing) have taken place, text is read and transformed into objects, which are then used to produce graphs, which, in turn, are used to produce a representation suitable for a Machine Learning algorithm.

### 3.3.2 From text files to complex sentence objects

This section will cover input processing for AlvisRE and how a text file is transformed to a set of complex objects. An object in computer science contains (structured) data and methods in a way that hides the data behind abstractions, and exposes functions that operate on that data [Martin, 2008].

An appropriately formatted text file can serve as efficient input to an IE system, transmitting all the complex necessary linguistic information as well as any existing annotations (NEs, relations, etc). The format used by AlvisRE, as well as the BioNLP format on which it is based, are standoff formats. A standoff format, as opposed to an inline format, is a markup (annotation) format where the annotations are given separately than the original text to which they refer. Similar formats are used to describe text in the BioNLP shared tasks (*e.g.* the BioNLP11 file format[17], the BioNLP09 GENIA task[18]), the format for Gene Regulation Event Corpus (GREC) [19], the one used by the popular annotation editor brat [Stenetorp et al., 2012][20], etc.

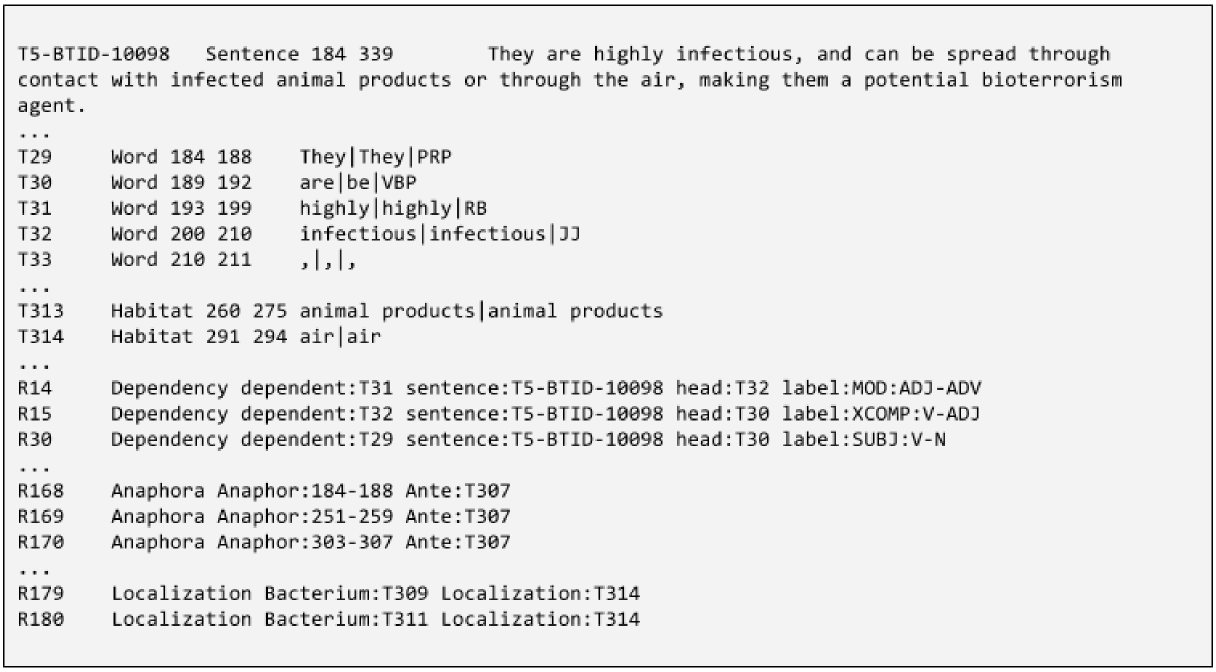
Three general types of annotations are found in formats of this type:

[17] http://2011.bionlp-st.org/home/file-formats

[18] http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/detail.shtml

[19] http://www.biomedcentral.com/content/supplementary/1471-2105-10-349-s1/standoff.html

[20] http://brat.nlplab.org/standoff.html



```
T5-BTID-10098   Sentence 184 339        They are highly infectious, and can be spread through
contact with infected animal products or through the air, making them a potential bioterrorism
agent.
...
T29     Word 184 188    They|They|PRP
T30     Word 189 192    are|be|VBP
T31     Word 193 199    highly|highly|RB
T32     Word 200 210    infectious|infectious|JJ
T33     Word 210 211    ,|,|,
...
T313    Habitat 260 275 animal products|animal products
T314    Habitat 291 294 air|air
...
R14     Dependency dependent:T31 sentence:T5-BTID-10098 head:T32 label:MOD:ADJ-ADV
R15     Dependency dependent:T32 sentence:T5-BTID-10098 head:T30 label:XCOMP:V-ADJ
R30     Dependency dependent:T29 sentence:T5-BTID-10098 head:T30 label:SUBJ:V-N
...
R168    Anaphora Anaphor:184-188 Ante:T307
R169    Anaphora Anaphor:251-259 Ante:T307
R170    Anaphora Anaphor:303-307 Ante:T307
...
R179    Localization Bacterium:T309 Localization:T314
R180    Localization Bacterium:T311 Localization:T314
```

Figure 3.1: Excerpt of an input file from the BB corpus, including a sentence (T5-BTID-10098), words (T29-33) figuring surface and canonical forms and POS tags, Named Entites (T313-4) with surface and canonical forms, dependencies (R14-5, R30), coreferences (R168-70) and semantic relations (R179-180). The complete file can be found in Appendix C.

- Text-bound, for words or parts of text, named entities and other singular entities or objects
- Relational, which declare a relation between other annotations
- Properties, which specify extra attributes for other annotations

In NLP literature different levels of linguistic annotations are often called layers. Buitelaar [Buitelaar, 2009] calls this layered NLP information "the NLP layer cake" and NLP software such as the AlvisNLP pipeline contain annotations that are often conceptually or programmatically organized in layers. Different layers define a theme or type (like semantic, syntax, grammar, canonical, ontology etc) and parts of text can be attributed annotations from different layers. AlvisRE follows a similar approach both in the format that it uses and in the object representations.

### The file format

Each annotation has an identifier unique to each input file. Only text-bound and relation annotations are defined in the input format of AlvisRE. Extra attributes on text are expected to be based either on words or named entities, and are listed together with the declaration of the corresponding text-bound word or named entity annotation.

#### AlvisRE object architecture

A set of input files is called a corpus. NE and semantic relation *types* are defined by the *schema,* which is given as an input and which is the same for the whole *corpus.* Each input file is a document, so a corpus is a set of *documents* with a given schema. The segmentation of the text of each document into *sentences* is given. So each *document* is represented as a set of *sentences.*

A sentence has *words*, *named entities* and *relations* associated with it.

Words are parts of text and optionally bear *tags* on defined layers (POS, semantic, etc). They are the result of tokenization and may, in reality, be words (linguistic sense), punctuation or multiple words in the particular case of them being multi-word expressions merged together in preprocessing.

Named Entities are parts of text that correspond to a set of words. Each NE object has a set of *words* attached to it. NEs can also have extra *tags.* They can be arguments of semantic relations, but this is not a requirement.

Three types of relations are possible:

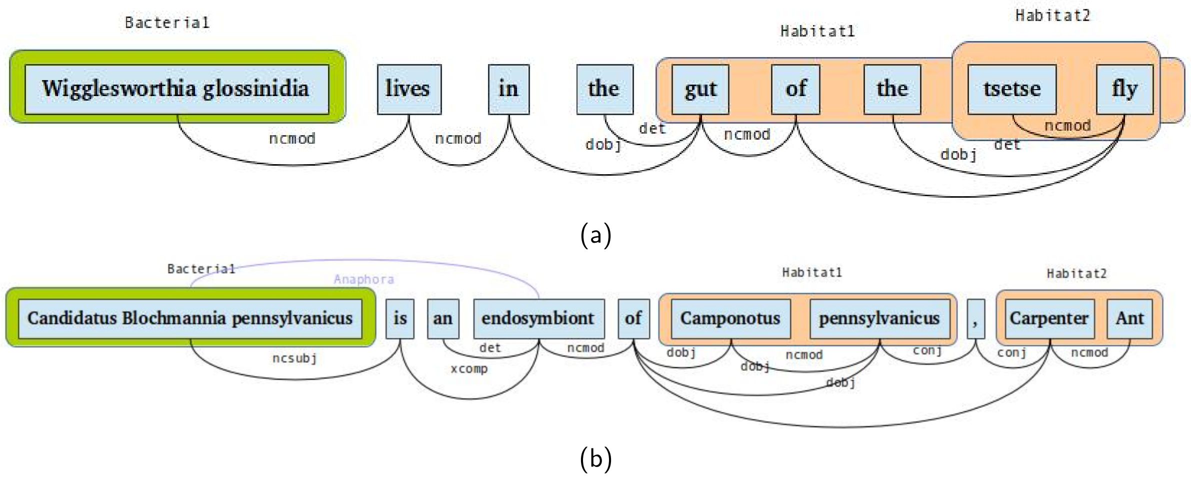
1. *Semantic Relations*, between *NEs.* They have a type and arguments (which are always NEs) in specific roles, according to a schema. They are either given in the input (for labeled corpora) or predicted by AlvisRE.
2. *Syntactic Relations*, between *words.* These are always syntactic dependency relations. They have a type, or label, which can vary depending on the syntactic parser used and they have arguments (words) in specific roles (generally head and dependent).
3. *Coreference relations*, a specific type of semantic relation. Their arguments are one antecedent, which is a *NE* and one anaphor, which is a part of text (*ie* a set of *words*). For a coreference relation, AlvisRE adds the anaphor *words* to the set of *words* attached to the referent *NE.* This allows AlvisRE to follow coreference links in the text.

### 3.3.3 From sentences to candidate relations

The annotation schema includes the definitions of valid semantic relations and defines what argument type combinations are possible, and the type of relation. A *candidate relation* or *candidate* can be defined as a co-occurrence of two possible arguments of a given type within a sentence. This co-occurrence of candidate arguments in a sentence is also the simplest approach for relation extraction and as such, it constitutes a baseline for the evaluation of RE systems.

Figures 3.2a & b show four candidate relations of type “Localization”, which take as arguments Named Entities of type Bacteria for the role of Bacteria, and Habitat for the role of habitat. In these examples, each sentence hosts two candidate relations, between the Bacteria entity and the two Habitat entities.

The set of the possible candidates together with their expected class defines the learning dataset. Candidates whose class is one of the defined relations are the positive instances, and candidates whose class is “no relation” are the negative ones. Depending on the choice of representation for



(a)

(b)

Figure 3.2: These two sentences come from the Bacteria Biotopes corpus and in them we can observe two types of NEs, Bacteria and Habitat. Bacteria multi-word entities have been merged to one token. Finally, a coreference relation is also illustrated in figure b.

the following steps, some candidate couples might be filtered out before learning and not used as learning examples. This occurs in cases of representations which require that arguments being connected in the syntactic parse graph. Although syntactic graphs should always be connected, dependency parsers can output disconnected graphs. This issue is discussed later on in sections 3.3.4 and 3.6.2.

#### Calculating candidates by co-occurrence

For each sentence, the subset of the cross-product of its NEs such that each pair of NEs is licensed by the annotation schema is added to the candidate list. A candidate has the same attributes as a semantic relation: arguments (NEs) in specific roles and a type.

A given argument $\mathrm{arg}_x$ composed of words ($w_{x1}$, .. $w_{xn}$) is considered to occur in a sentence s if and only if any of its words $w_{xi}$ occurs in s in this order, or if there exists an anaphor $a_j$ that occurs in s and refers to $\mathrm{arg}_x$.

For each possible pair of arguments $\mathrm{arg}_x$ and $\mathrm{arg}_y$ which occur in a sentence and can be in a relation of type $t_i$ according to the schema, occupying roles $r_{i1}$ and $r_{i2}$ respectively, a candidate relation $c(t_i, r_{i1}(\mathrm{arg}_x), r_{i2}(\mathrm{arg}_y))$ is added to the candidate list for this sentence.

### 3.3.4 From candidate relations to paths

#### Introduction

Sentences translate naturally into graphs with nodes being words and edges being relations -syntactic or other. Graphs have thus become a popular choice for semantic relation representation. A good representation of the candidate relations includes as much available information as possible, and graphs allow for this sort of flexibility in representation.

Paths are the basis of this work as a representation. A path is a sequence of nodes (N) and edges (E), which starts with a node, alternates nodes and edges, and ends with a node. For example the following three sequences could be valid paths of lengths 5, 17, and 3 respectively:

- $N_1$ $E_1$ $N_2$ $E_2$ $N_3$
- $N_1$ $E_1$ $N_2$ $E_2$ $N_3$ $E_3$ $N_4$ $E_4$ $N_5$ $E_5$ $N_6$ $E_6$ $N_7$ $E_7$ $N_8$ $E_8$ $N_9$
- $N_1$ $E_1$ $N_2$

In the case of AlvisRE, nodes are words and edges are linguistic relations. The term linguistic relations is used here to group syntactic relations such as dependencies and relations of neighbourhood between words.

Two types of graphs are used: the dependency graph and a graph where each word is connected with its neighbours by word order. AlvisRE uses a shortest-path algorithm on each type of graph to find a path between the two arguments of a candidate relation. Based on these two types of graphs, three types of paths are used:

1) Dependency paths: paths on the dependency tree of the sentence containing a candidate, computed using Dijkstra's single-source shortest path algorithm, which works with undirected graphs and has a time complexity of O(E + V log V), when using binary heaps [Barbehenn, 1998] as AlvisRE does.

2) Word-order paths defined as subsequences of the sentence following the order of the words in the sentence. In this particular case, the paths follow the order of words in the sentence and a shortest-path algorithm is unnecessary.

3) A third type which combines the other two is possible, proposed as a solution to incomplete parse trees. Dijkstra's algorithm also works with weighted edges, a property that AlvisRE can use to prioritize certain types of relations -in this case syntactic over word-order relations.

### Dependency Paths

Dependency grammars are used to describe the syntax of a sentence and are based on words and dependency relations between them. Dependency relations are labeled and directed binary relations between two words. They take two arguments, called head and dependent.

Some examples of dependencies are

- between verb and noun, e.g. subject-of object-of
- between modifier and modified, for example adjective and noun
- between preposition or determiner and noun
- coordination relations, for example when two nouns are separated by "and", or commas.

Dependency parsers are programs that produce data structures, such as graphs or trees in particular, based on a dependency grammar. Ideally, the parsing of one sentence results in a single connected graph. While this is true for manual syntactic annotation, dependency parsers can often produce imperfect parses, resulting in disconnected subgraphs. Notable dependency parsers used in RE are the Stanford Parser [De Marneffe et al., 2006], Enju [Miyao et al., 2003] and, specifically for the biomedical domain, the CCG parser [Clark and Curran, 2007].

In most cases dependency graphs are trees, and as such acyclic,which results in the existence of a unique path between any two words. In some grammars co-ordination might create cycles, in which case the dependency graph is not a tree. Even though dependencies are asymmetric relations and define a directed graph, when calculating paths between words for the task of RE, the direction of the relations is not taken into account in AlvisRE.

In RE dependency graphs and paths in particular are common representations. Bunescu and Mooney [Bunescu and Mooney, 2005] introduce the “shortest path hypothesis”, which states that if two entities are in relation, the most relevant part of the dependency graph to this relation is found almost exclusively on the shortest path between the two entities, even if the graph is not acyclic due to coordination.

### Dependency Paths in AlvisRE

For each candidate relation, AlvisRE uses Dijkstra’s algorithm to calculate the path connecting the two arguments on the dependency tree of the sentence this candidate relation belongs to. This path includes the arguments.

For each sentence s, having a dependency tree G, composed of words $\mathrm{w_i}$ as nodes and dependencies $\mathrm{d_j}$ as edges :

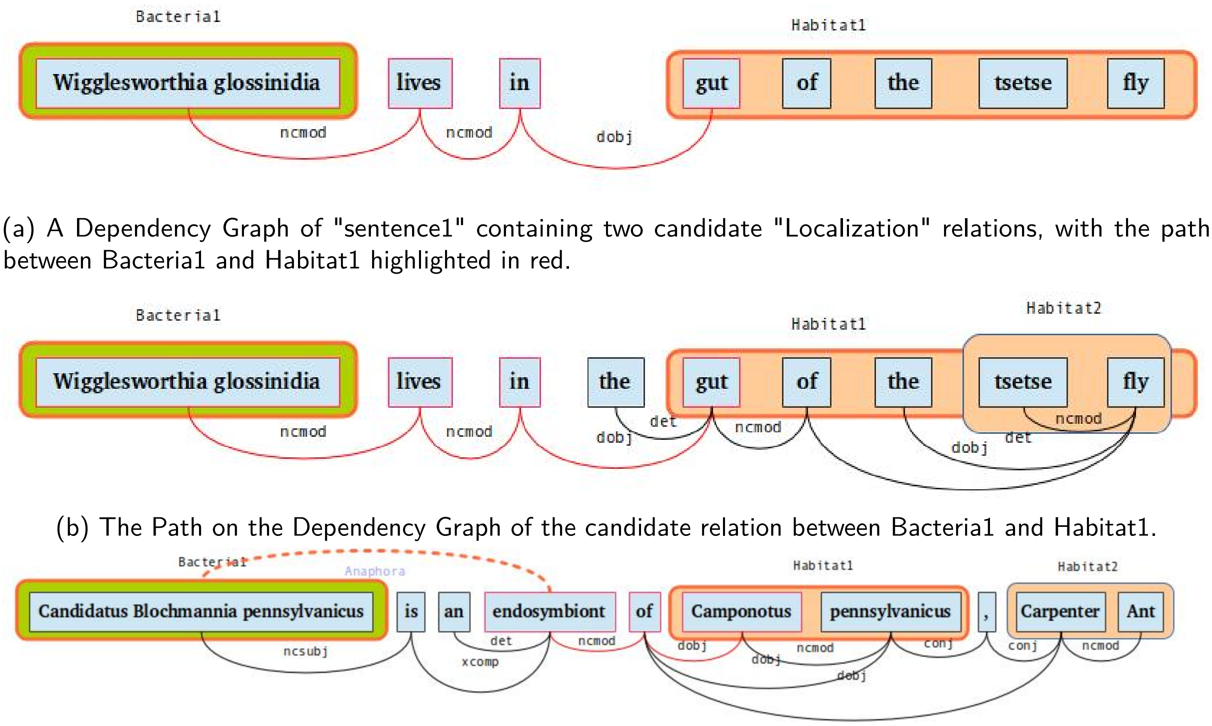
- For each candidate relation *c* with arguments $arg_x$ and $arg_y$, belonging to sentence *s*.
- if $W_x$ and $W_j$ are the sets of words of $arg_x$ and $arg_y$
- Dijkstra’s algorithm is used to find all possible paths between $W_x$ and $W_j$ on $G$, and keep the shortest one.

### Multi-word entities and shortest paths

Multi-word entities are, simply, Named Entities which contain more than one word. In some approaches multi-word entities are grouped together (ex. [Reichartz et al., 2010, Yakushiji et al., 2006, Miyao et al., 2009], or replaced by a single word (ex, Buyko 2011). In other approaches only the heads of multi-word entities are kept in a simplified minimal subtree (ex. [Chowdhury and Lavelli, 2011]).

If after the preprocessing step multi-word entities remain, AlvisRE keeps the shortest possible path containing the head word of the argument entity. This idea is similar in the spirit as the simplification of trees, as they are based on the hypothesis that in multi-word entities, it is the head of such an entity that will always be connected to the rest of the tree. Figures 3.3a, 3.3b and 3.3c include examples of this phenomenon.

Multi-word entities naturally occur in corpora. Additionally, the way AlvisRE treats coreference also produces multi-word entities, with antecedents being linked to anaphors. If both an anaphor and an antecedent occur within a sentence, AlvisRE still chooses the shortest path possible, as a convention and in the spirit of [Bunescu and Mooney, 2005] by keeping the most concise representative subsequence available. (see Figure 3.3 for an example).



(a) A Dependency Graph of "sentence1" containing two candidate "Localization" relations, with the path between Bacteria1 and Habitat1 highlighted in red.

(b) The Path on the Dependency Graph of the candidate relation between Bacteria1 and Habitat1.

(c) The Dependency Graph with the Path of the candidate "Localization" relation between Bacteria1 and Habitat1, on "sentence2", highlighted in red.

Figure 3.3: Paths from the Bacteria Biotopes corpus, which includes coreferences. Bacteria multi-word entities have been merged together, but habitat entities can include multiple words.

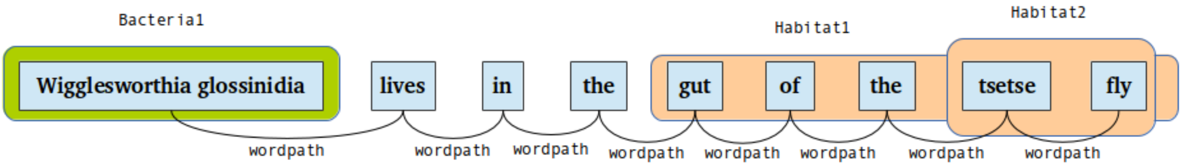


Figure 3.4: "Entire Sentence" representation: in the same sentence seen previously as "sentence1", the path between the first and last words is used.

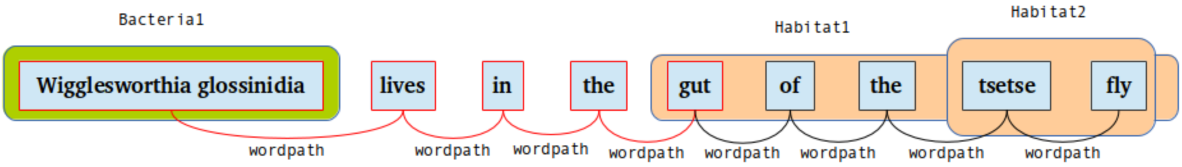


Figure 3.5: "Surface Path" representation: the path between the two candidate arguments (in red) is used.

### Word-Order Paths

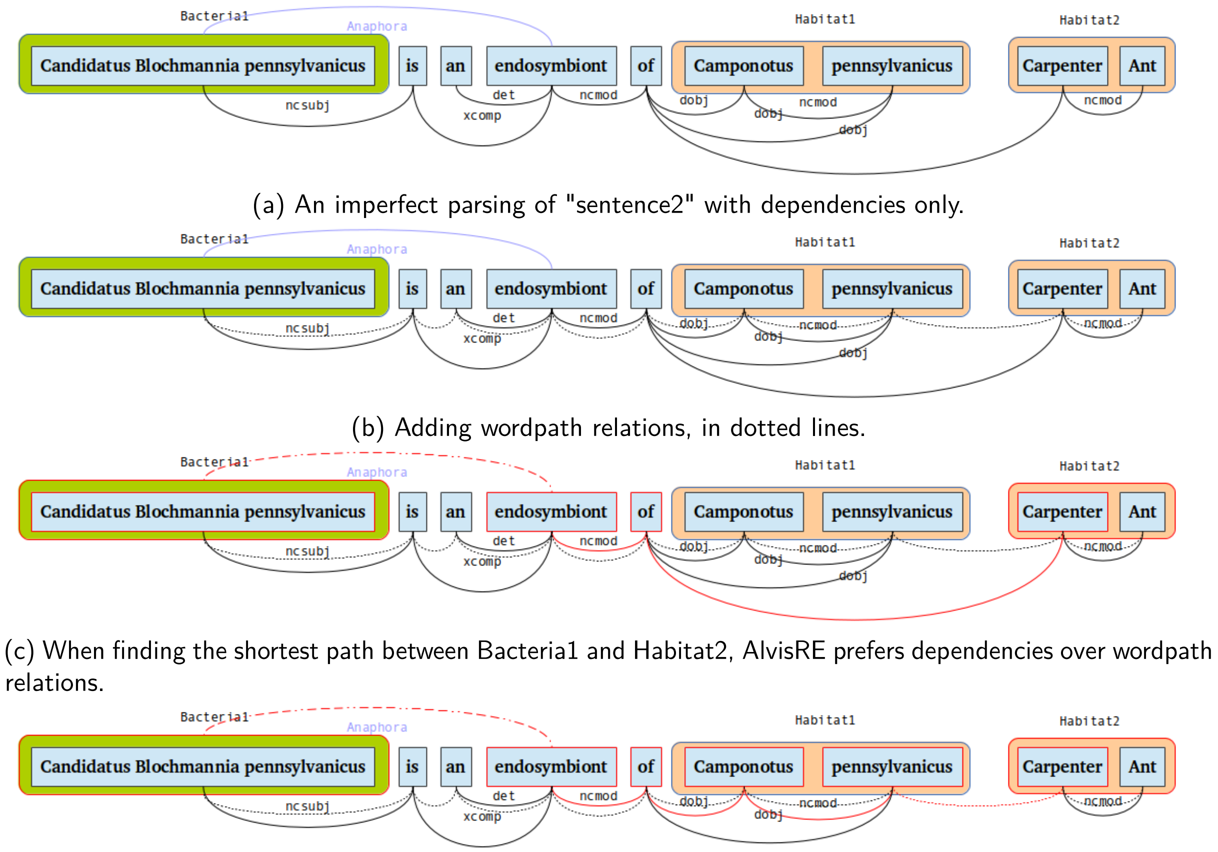
Word-order paths follow the order of the sentence and not the syntactic dependencies. They were introduced into AlvisRE in order to be able to compare and evaluate the choice of dependency paths. In order to have a homogeneous representation of the paths, AlvisRE uses a relation called “wordpath”. It is represented as a neighborhood relation with the label “wordpath” and two arguments with identical roles. These relations are produced during preprocessing.

Word-order paths are closer to the notion of co-occurrence of arguments within a sentence and they are not influenced by the performance of dependency parsers. However, in argument co-occurrence the context of the arguments is not represented at all, making word-order paths a potentially more powerful tool. Additionally to the comparative evaluation of dependency paths, this representation is also particularly useful for testing the co-occurrence scenarios and building a baseline.

Two types of word-order path representations are used by AlvisRE. The first is the path between arguments, called here a “Surface Path”. This type of path is defined as the sequence of the words between the candidate arguments (inclusive), in the order they appear in the sentence. The second one is the entire sentence as a path. This representation equals the path between first and last words (inclusive) of the sentence. With this representation, when a candidate relation is detected in a sentence, the text of the entire sentence is used as a representation of this candidate relation.

### Combined Paths

A combined path is a path containing both dependencies and wordpath relations. They were introduced in order to counter the problems of imperfect dependency parsings in the form of disconnected graphs. As pointed out above, no learning example can be derived since the



(a) An imperfect parsing of "sentence2" with dependencies only.

(b) Adding wordpath relations, in dotted lines.

(c) When finding the shortest path between Bacteria1 and Habitat2, AlvisRE prefers dependencies over wordpath relations.

(d) IF Habitat2 was completely disconnected in the dependency graph, the wordpath relation connecting to the rest would be selected.

Figure 3.6: A modified version of "sentence2" used as an example of parsing problems and fixing them with worpath links.

candidate arguments reside in different subgraphs and are not linkable by a path. Such parsing problems do not appear to follow a pattern or have any predictive regularity, but appear to be random, instead [Ratkovic, 2014]. This makes targeted reparations very hard to engineer.

By combining wordpath links with dependencies, disconnected dependency graphs become connected. The focus of our approach remains in using dependency paths and limiting the use of wordpath links, so weights are introduced on the edges to favour syntactic over wordpath links. The shortest-path algorithm is discouraged from using wordpath edges by penalizing them with a high cost (100), when dependencies have a cost of 1. This ensures that whenever a path using only dependency edges exists, it will be chosen and that wordpath edges will exclusively be used to connect disconnected parts of the graph.

In the Figures 3.6a-d *sentence2* of the previous examples has been modified[21] to show possible parsing problems and how combined paths can repair the phenomenon.

[21] the elements removed were chosen to facilitate the visual representation and do not necessarily reflect actual parsing errors.

### 3.3.5 From paths to a machine-learning ready representation

#### Introduction

The next step requires a transformation into a representation that can be directly exploited by a machine learning system. As was discussed in Chapter II, RE systems use both feature and kernel-based approaches. While feature-based approaches often use all types of information, kernel-based approaches to RE use in their vast majority kernels based on syntactic trees, with graph kernels being the most commonly used kernels for RE.

AlvisRE can work with different types of representations (paths), as explained in the previous section, and can use either co-occurrence or machine learning for semantic relation extraction. The machine learning approach used by AlvisRE is based on a sequence alignment method, called Global Alignment. This method aligns sequences based on their similarity and calculates that similarity while doing it. It is used as a similarity function in AlvisRE.

Global Alignment is a method created for sequence alignment of biological sequences, but it was adapted to work on the path representation described in the previous section. The similarity matrix calculated using Global Alignment is then transformed into a vectorial representation by a method called Empirical Kernel Map.

#### Co-occurrence

Co-occurrence is the occurrence within one sentence of two NEs which can be arguments of a relation according to the schema. It is a necessary condition for relation extraction, but in the simplest approaches it can also be considered a sufficient condition. A slightly modified approach is based on the condition of co-occurrence of arguments and trigger words within a sentence [Ratkovic et al., 2012].

For all the different types of representations that AlvisRE supports, co-occurrence is used as a necessary condition (see candidate relations above). It is also used as a baseline approach in testing. Co-occurrence with trigger words is also implemented. The co-occurrence of NEs and possible trigger words is defined for each case of representation:

- On the dependency path
- On the surface path (arguments and words between them)
- In the entire sentence (typical scenario).

#### Paths as sequences

We can consider a path as a sequence of nodes and edges. Consequently, the change of representation from paths to sequences is not a transformation, but a rewriting. Instead of a graph with relation edges between words, a path is rewritten as a sequence of basic elements, which can be either words or relations (see Fig. 3.7). Essentially, this representation is a sequence view of a path and is in the form of a sequence of words and dependencies alternating, starting and ending with a word.

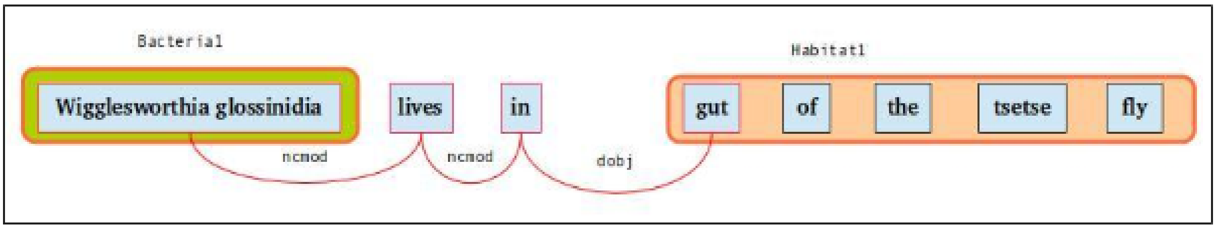


Figure 3.7: If the words "W. glossinidia", "lives", "in" and "gut" have the IDs $w_1$, $w_2$, $w_3$, and $w_4$ respectively. And $r_1$ = ncmod($w_1$, $w_2$), $r_2$= ncmod($w_2$, $w_3$), $r_3$ = dobj($w_3$, $w_4$). Then the candidate of sentence1 highlighted in this figure can be rewritten as: ($w_1$,$r_1$,$w_2$,$r_2$,$w_3$,$r_3$,$w_4$).

Written as a regular expression, this representation would be (wd)+w. In more detail, for arguments $\text{arg}_\text{x}$ and $\text{arg}_\text{j}$, the first word of the sequence belongs to $\text{W}_\text{x}$ (the set of words of $\text{arg}_\text{x}$) and the ending word to $\text{W}_\text{j}$. Of the possible words belonging to the argument entities, these two are the pair of words that was used to obtain the shortest path. An example can be seen in Figure 3.7.

### Global Alignment and a similarity-based representation

AlvisRE uses a similarity function called global alignment. Global alignment (GA) algorithms align entire sequences by calculating the similarity of their parts. They are mainly used in alignment of biological sequences. *Global*, as opposed to *local* alignment techniques, put more weight in aligning the entire sequence. The algorithm used in this work is a dynamic programming algorithm called the Needleman-Wunsch algorithm [Needleman and Wunsch, 1970].

Dynamic programming algorithms solve complex problems by splitting them into simpler and easier to solve subproblems, and making sure that each subproblem is solved just once, typically by memorizing their solutions. In the case of sequence alignment, aligning a sequence is achieved by calculating the cost of aligning their parts by computing their similarity.

The Global Alignment algorithm calculates the similarity between a pair of path sequences (p1 and p2) by the following function

$$F_{ij} = \max_{h<i,k<j} \{F_{h,j-1} + S(A_i, B_j), F_{i-1,k} + S(A_i, B_j)\}$$

where $A_i$ are the elements of p1, and $B_j$ are the elements of p2.

### The alignment algorithm

- Let *ppscore* be the similarity function between two elements, or point-to-point score (detailed below).
- F is the matrix which includes the optimal alignment score between each pair of elements, $\text{p}_1$ and $\text{p}_2$. It consists of the scores $\text{F}_\text{ij}$ where i is the index of an element belonging to $\text{p}_1$ and j is the index of an element belonging to $\text{p}_2$.
- The cost of aligning an element with a gap is called *gapPenalty*.
- $\text{F}_\text{ij}$ is calculated by choosing the maximum of these scores:

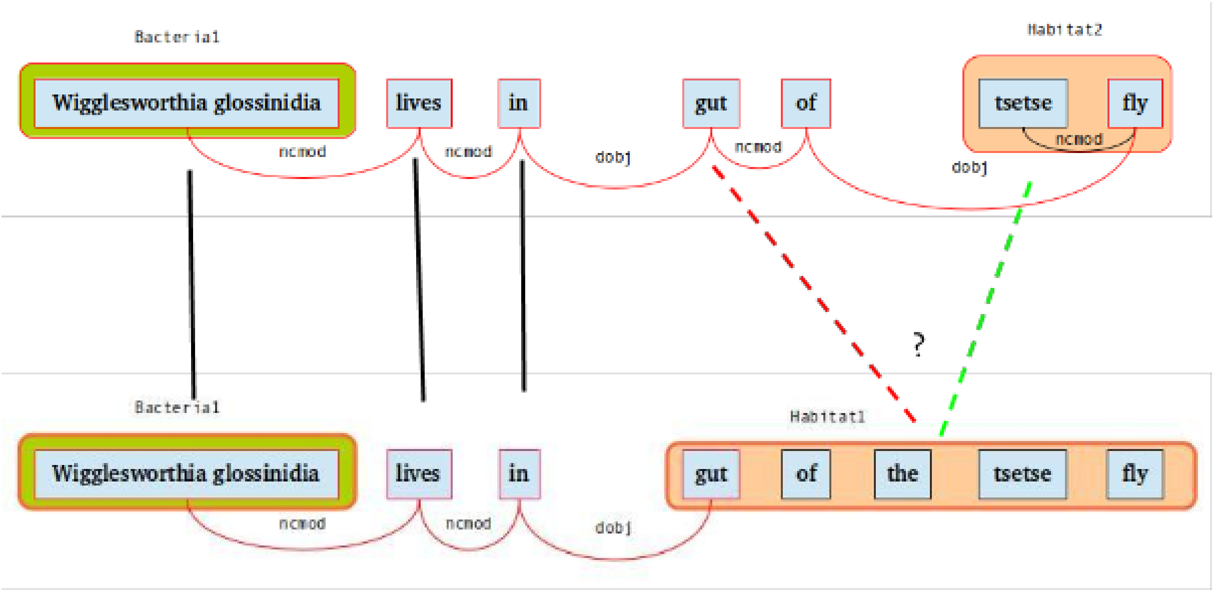


Figure 3.8: Gap penalty should be such that, in "sentence1", if the two possible candidates "Bacteria1 - Habitat1" and "Bacteria1 - Habitat2" were aligned, "Habitat1" would ideally be aligned with "Habitat2" and not the word "gut".

- $F_{i\text{-}1,j} + gapPenalty$
- $F_{i,\ j\text{-}1} + gapPenalty$
- $F_{i\ -\ 1,j\ -\ 1} + ppscore(\text{i,j})$

- The optimal alignment score between p1 and p2 is found in position $F_{mn}$ and is normalized for the length of the paths as follows:

$$\bar{F}_{mn} = \frac{2*\ F_{mn}}{m+n}$$

- *ppscore* takes values in {0,1} and thanks to this normalization (and a *gapPenalty* also in {0,1}), the alignment score of two sequences F takes values in {0,1} too.

The complexity of this algorithm is O(mn) in both time and space, where m and n are the lengths of the two sequences to align.

The default value of *gapPenalty* is 0.3, a value which was empirically found to give the best results in testing. The idea behind such a value is to not penalize the existence of extra information or longer expressions and to cope with some linguistic phenomena like ellipsis, inserts, *etc.* All default parameters were calculated by cross validation on the LLL and BioNLP-ST ’13 BB corpora. Figure 3.8 shows an example that illustrates this principle.

### The scoring function *ppscore.*

A number of parameters provided by the user of AlvisRE can tweak the behaviour of *ppscore* (and the *gapPenalty*). Only the default values are given in this section, along with an interpretation. The alignment of words to relations is not possible, thanks to a fixed very low score assigned to such an alignment.

In calculating the similarity between two **relations** $r_1$ and $r_2$, if $r_1$ and $r_2$ have the same label, ppscore returns by default the maximum score, 1.0. In case of different labels, three distinctive cases can occur: a) a similarity table between syntactic labels is used if one was provided by the user, or b) one of them is a wordpath relation, in which case a default score of 0.7 is returned, or, c) otherwise a default "dissimilarity score" of 0.2 is returned.

Since the wordpath label is used when the parser has failed to produce a complete parsing, the alignment of dependencies is preferred to an alignment of two wordpath and syntactic dependency labels.

The syntactic label similarity table can prove useful in the case of dependency grammars where some syntactic labels play related roles, defining syntactic families. An example is with the Alvis Grammar parser complement labels "COMP_of:V-N" and "COMP_with:V-N". In this example, even if the preposition is different it is wiser to not consider them completely dissimilar.

The similarity between **words** $t_1$ and $t_2$ is calculated as the weighted average of the similarities in each layer = $\frac{s_i w_i}{w_i}$ , where $s_i$ is the similarity between the labels (values) of the two words for layer *i,* and $w_i$ is the weight of layer *i.* These weights are chosen in a way that the range of this similarity remains [0,1].

### Weights

For the three standard layers of part-of-speech (POS) tags, surface form, and canonical representation, the default weights are 0.1, 0.5 and 1.0 respectively. A weight of 0.1 for the POS layer means that a similarity of grammatical function is considered as slightly better than complete dissimilarity. While semantic objects rarely occur having different grammatical roles, it is still useful to prioritize the alignment of two otherwise dissimilar words having the same role over two having a different grammatical role.

Surface similarity is string similarity and it is considered an even better indicator of semantic similarity than POS. It has a lower weight than canonical similarity so that it does not penalize semantic objects that are identical but written in a different manner, for example "B. subtilis" and "Bacillus subtilis (strain 168)", in the case of the bacteria entities in the BB corpus.

Canonical similarity compares lemmatized or normalized forms and is considered the best indicator of semantic similarity. The "B. subtilis" examples above have the same canonical form "Bacillus subtilis".

Finally, for any additional layer defined, a weight needs to be provided by the user.

*Layer-specific similarity functions*

For the POS layer and any user-defined layer, the similarity of the labels is calculated either by the use of a user-provided similarity table, or if there is none a default binary similarity function is used, which calculates the similarity score as follows:

- $s_{POS}$= 1, if the labels are identical
- $s_{POS}$ = 0, otherwise

For the surface layer, a string similarity based on the Sørensen–Dice coefficient is used [Sørensen, 1948, Dice, 1945]. Other alternatives are possible and listed in section 3.5.4 of this chapter.

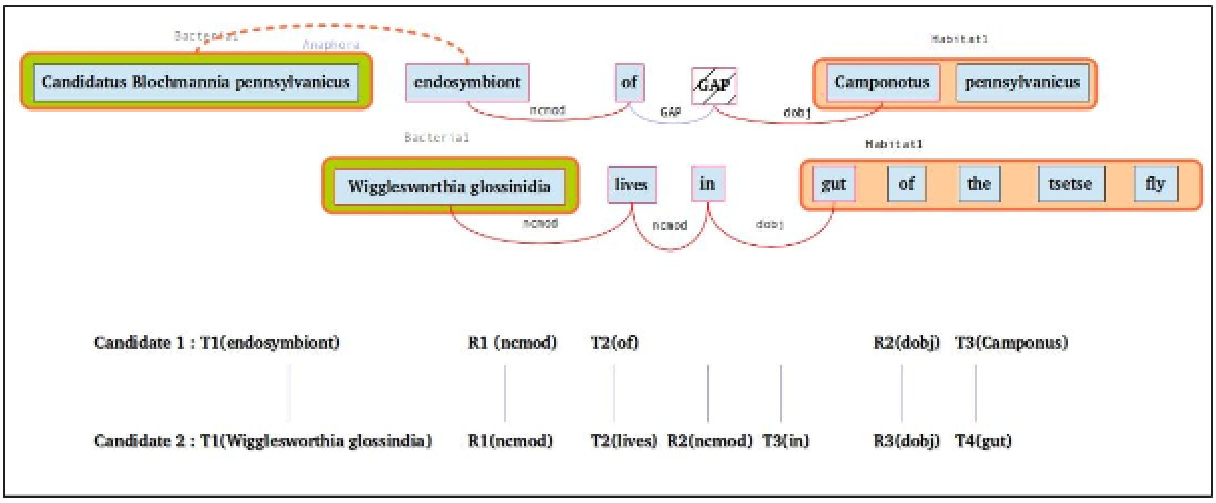


Figure 3.9: The alignment of two candidates. The second example includes an error in parsing, where *lives* has been parsed as a noun instead of a verb, and the syntactic relations linking it to its neighbours are not those of a verb object and subject. In spite of this error, AlvisRE manages to align those two candidates in a correct way.

For the canonical layer, if the words have the same canonical form, $s_{\text{Canonical}} = 1$, otherwise, a number of tools for computing the semantic similarity can be used. For more details see Section 5 of this chapter. Additionally, if multiple semantic lookup tools are enabled, the best score is kept. If the list of their answers is $A \in \{0, 1\}$, $s_{\text{Canonical}} = \max\ (A)$. This choice was made based on the nature of the semantic tools. Finally, if no canonical layer information has been provided, the semantic similarity method is computed from the surface representations of the words.

Note that it is optimal to give surface similarity a weight of 0 for a corpus where good quality lemmatization and normalization has taken place.

### Configuration parameters

A detailed documentation of the configuration parameters for the GA (Global Alignment) similarity function can be found in Appendix G. Additionally, the section "Experiments" contains details about the sets of experiments performed to evaluate the influence of selected parameters and determine their default values.

### Example

An example of alignment between relation candidates using this method can be seen in Figure 3.9. This figure shows the alignment of the candidates seen in previous examples, in path and sequence forms.

### Vector transformation

Similarity functions are also called kernel functions. The necessary condition for a similarity function to be used as a kernel for Machine Learning is called Mercer's condition, and it states that only similarity matrices that are positive semi-definite (PSD) are safe to be used with a kernel method [Minh et al., 2006, Schölkopf and Smola, 2002]. The Global Alignment kernel function does not possess these mathematical properties as the resulting matrix is not a positive semi-definite matrix. A transformation called Empirical Kernel Map (EKM) [Tsuda, 1999, Schölkopf and Smola, 2002, Mingrui Wu et al., 2006] is used instead. Using EKMs is a good solution for global alignment methods as it produces a valid kernel matrix, but such an approach additionally makes it possible to incorporate prior knowledge into the kernel [Liao and Noble, 2002, Liao and Noble, 2003, Schölkopf et al., 2004]. In our case the prior knowledge is all the different similarity scores used to produce this alignment, which allow us to use elaborate knowledge, like semantic information for example.

The EKM transformation produces a feature vector for each learning instance. The EKM produces a feature vector built on a similarity function and avoids feature engineering, but can be combined with additional features optionally. This new vectorial representation can then be used with standard kernels, such as the Euclidean kernel.

For each learning instance $x_i$, a vector ${}_{x}$ is created by the EKM transformation. This vector ${}_{x}$ is defined as follows:

- Let $\{l_1, \ldots, l_n\}$ be the set of labeled instances, and
- $s_{ij}$ be the similarity of instance $x_i$ to labeled instance $l_j$
- ${}_{x} = \{s_{i1}, \ldots, s_{in}\}$

This transformation means simply that each learning instance is represented by a vector of the similarities between this instance and all training instances. For the $s_{ij}$ similarity, the Global Alignment similarity is used. For the training (labeled) set this map equals the symmetric similarity matrix for all instances of the set. This final form of representation is the one used for classification in the next step. An initial training set for each learned model is necessary for any future classification task. The use of the EKM method on the symmetric training similarity matrix is described in Figure 3.10.

#### Using varied sources of information

AlvisRE has been designed in a way that allows using different sources of information in different ways. The path representation can include many layers of information, such as surface forms, canonical or lemmatized forms, grammar, syntax and semantic information such as semantic classes or ontology concept associations.

Additionally to this flexible representation, the Global Alignment similarity is based on the principle of the similarity of parts and can make full use of the multi-layer linguistic information. This design also can integrate external tools for providing similarity of parts such as textual or semantic similarity. For semantic similarity, semantic classes or other semantic information can be used, such as resources (e.g. WordNet, Ontologies), or computed by tools (e.g. word2vec) *etc.* (see section “Semantic Information” below).

The vector representation after the EKM transformation additionally allows for extra features, such as trigger words, any features typically used in RE approaches or other features for speculation/negation detection that are not taken into account in similarity computation but can serve as additional filters.

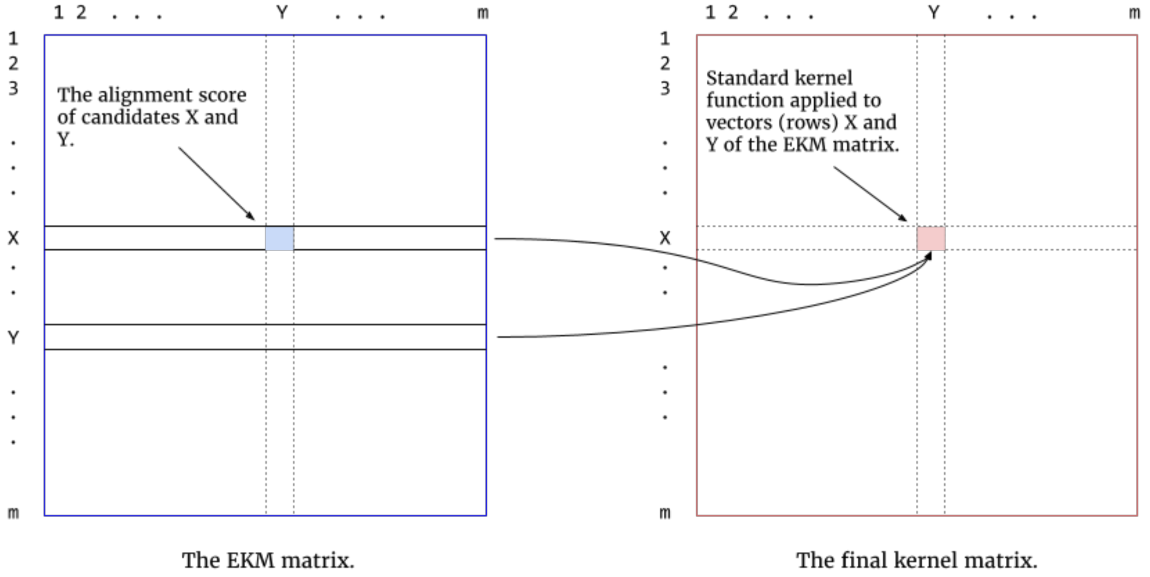

1 2 . . . Y . . . m
1
2
3
X
Y
m
The alignment score of candidates X and Y.
The EKM matrix.
Standard kernel function applied to vectors (rows) X and Y of the EKM matrix.
The final kernel matrix.

Figure 3.10: A closer look at the EKM transformation and resulting kernel matrix, in the case of the training dataset. In the EKM matrix each candidate is represented by a row composed of the alignment scores of this candidate with every other candidate. For example, for the candidate X, its representation is a vector which corresponds to row X of the EKM matrix, and where the column Y contains the Global Alignment score between X and Y. Later, during classification any standard kernel function can be used. This method will produce a kernel matrix on its own, which will use the vectors of the EKM matrix as input to compute the similarity of X and Y in a way that is guaranteed to satisfy Mercer's condition.

### Algorithmic complexity

The complexity of the calculation of the dependency paths is that of Dijkstra's algorithm, O (E+VlogV) where E is the number of words and V is the number of syntactic relations in the dependency graph. With an average of 26 words and 46 dependencies per sentence for the *Arabidopsis* corpus and a linearithmic complexity, the computation time for each candidate remains reasonable. Even in the case of large corpora, this results in minutes of computation time (CPU time).

The complexity of the Global Alignment algorithm is O(mn) where m and n are the lengths of the paths to be aligned. This means that all the calculations necessary for each point-to-point score will be performed m*n times in order to obtain the alignment score of a couple of candidates.

The complexity of the EKM transformation is O(m'n'), where m' is the dimensionality of the vector, *i.e.* the number of training candidates, and n' the number of candidates for which the vectorial representation is calculated. Consequently it is O(m'$^2$) for the training set. This translates to m'*n' times of running the Global Alignment function. While this complexity remains polynomial, AlvisRE using Global Alignment can be quite demanding in computation time. For example, for a relatively small corpus with 1000 candidates of average path[22] length of 20, for the training set this translates to:

- *ppscore* will run ~400 times for the alignment of each pair of candidates.
- The transformation will calculate 500.000 (since it is symmetrical) alignment scores.
- *ppscore* will run in total 200 million times

Consequently if *ppscore* takes 1ms each time this means it takes ~56 hours of computation time to get the input for the classifier. AlvisRE has been implemented using multi-threading whenever possible –and namely during the EKM transformation, as well as other optimizations and caching, but the complexity of the algorithm should be taken into account when dealing with large datasets. Rich representations using external semantic tools as the latter can add a serious overhead, leading to week-long computations.

## 3.4 Classification

### 3.4.1 Introduction

The RE task can be defined as a classification (or supervised learning) task with the relation types (of the schema) being the labels to learn. An additional label "no relation" is used for the negative instances, or instances that do not belong to any of the relations defined by the schema. Consequently, at prediction time, the ML algorithm assigns new instances either to a known-type class, or to the "no relation" class.

In most RE tasks more than one type of relations are possible and so RE is generally a multi-class classification problem. In the case of additional properties and modalities of relations, such as

[22] The average sentence in the Arabidopsis corpus has 26 words. Paths include both words and relations in the dependency path, so if we assume that about 1/3 of the words end up in the path, this results in paths of approx. average length of 20.

speculation, negation, etc, RE could also be viewed as a multi-label classification problem. In practice, this is often treated as a two-step classification problem.

The EKM transformation introduced in section 3.3.5 produces a vectorial representation that can optionally be enriched with additional features. These feature vectors can be used by a number of standard supervised learning algorithms. AlvisRE is compatible with all the algorithms implemented in the Weka (Waikato Environment for Knowledge Analysis) machine learning software [Hall et al., 2009].

Even though kernel methods and similarity functions are best understood intuitively in use with nearest-neighbours approaches, the best-performing algorithm through repeated experimentation was found to be SVMs, outperforming other methods by more than 15 points in F-Measure. The libSVM and libLINEAR implementations of SVMs have both been tested and used.

### 3.4.2 Support Vector Machines

Support Vector Machines (SVMs) are used primarily for classification, but can also used for regression. They perform classification tasks by constructing hyperplanes in multidimensional space that separate examples from different classes. To construct an optimal hyperplane a SVM employs an iterative training algorithm that minimizes an error function.

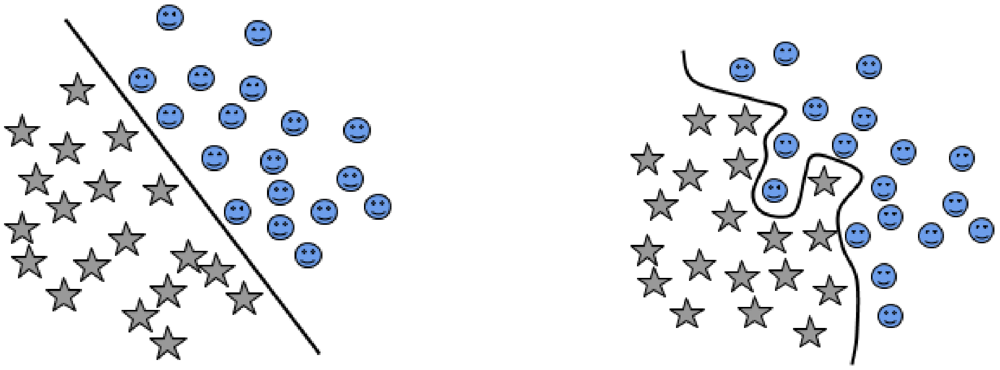
What the SVMs do is producing nonlinear boundaries by constructing a linear boundary in a large, transformed version of the feature space [Hastie et al., 2001a]. This quality is what makes SVMs ideal for data that are not linearly separable, such as the data for RE. By making a non-linear transformation of the original input into a high dimensional feature space, SVMs offer the possibility of better classification by finding an optimal separating hyperplane in this feature space. This transformation is actually performed by the kernel function, and this process can be described as the kernel trick (also mentioned in section 1.6.3.0). Figures 3.11a-b and 3.12 illustrate the difficulty of finding a separator for nonlinear spaces, and how kernel methods address this problem.

For the SVMs, an optimal separating hyperplane is defined by a maximal margin classifier based on the training data, the samples which lie on the margin are called support vectors (see Figure 3.13).

The soft margin parameter (or c parameter) of SVMs is a penalty factor and is the most important setting for optimizing the performance of a SVM.

- It is a regularization parameter, which helps with generalization.
- The c parameter controls the influence of misclassification, with larger c values leading to smaller margins for the separating SVM hyperplane.
- The tradeoff is between small margins, which mean fewer misclassifications in the training data, and larger margins with greater model complexity, which can avoid overfitting.

Multi-class classification with SVMs is performed as “1 vs all” classification. This strategy consists in training a single classifier per class, where examples of this class are considered positive, and all other examples are considered negative, that is to say all false candidates of this class and all positive and negative candidates of other classes. In the case of RE and the



(a) A linearly separable classification problem with a linear separator.

(b) A non-linear separation

Figure 3.11: Linear and non-linear separators for classification.

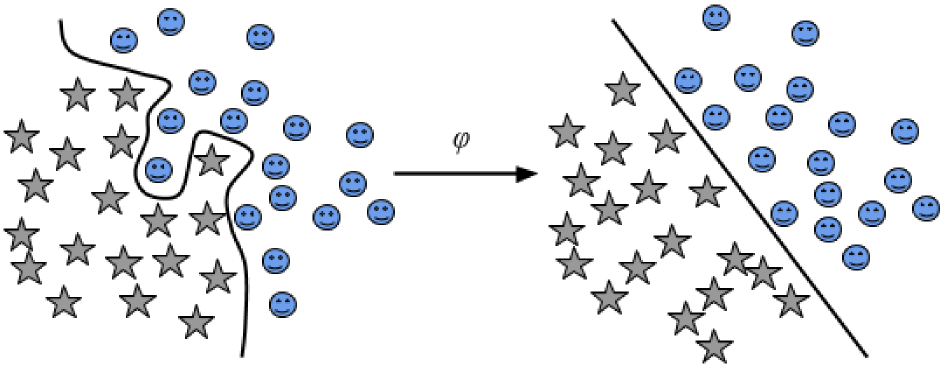


Figure 3.12: A non-linear input space transformed into a linear feature space by a kernel function $\phi$.

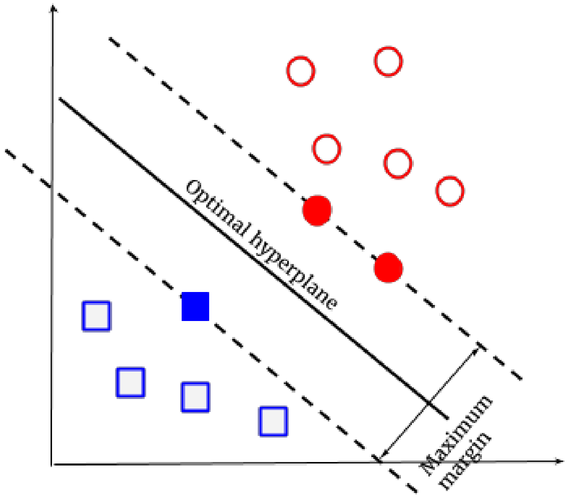


Figure 3.13: Maximum-margin hyperplane and margins for an SVM. Samples residing on the margin are called the support vectors.

representation and similarity function choices of AlvisRE, it might happen that different semantic relation types end up containing examples which might be close to example from other relation types. This can be a problem in the "1 vs all" scenario, as it creates a much more complex and noisy space. A different approach consists of skipping SVMs' multi-class classification and conducting separate classifications for each semantic relation type. In this "per class" strategy, the positives for each class are the candidate relations of this specific type that are indeed true relations, and the negatives are the candidate relations of this type, again, that are false.

Finally, SVMs allow for weights that can be used to balance relatively underrepresented classes (here, relations). By assigning weights to classes, errors in classification of classes with higher weights during training are penalized more, compared to classes with lower weights. In both strategies of "1 vs all" and per class classification, imbalances in representation of relations occur often in RE. In the first case, one relation might simply have fewer examples in the training set and negative examples (possible candidates not actually belonging to any class) are the majority. In the second case, for per class classification, it is almost always the case that negative candidate relations vastly outnumber positive ones.

In practice, the c parameter for the SVMs, as well as the ideal class weights are optimized during training.

## 3.5 Adding Semantic Information

### 3.5.1 Introduction

Semantic information is the information on sense and meaning. The goal of RE is to extract semantic relations between entities and it operates at the level of sentences or phrases. Traditional sources of semantic information are dictionaries, ontologies or databases. Additionally, distributional semantics is an approach based on the principle that parts of text that have similar distributions have also a similar meaning. Distributional semantics systems are statistical or machine learning systems, and rely on large volumes of text to calculate the distributions of words or phrases [Sahlgren, 2008]. They can use domain-specific corpora to produce rich resources that are better adapted to a specific RE task.

It is worth noting that the term "semantic information" as it is used here is not related to the "semantic relations" that are extracted by RE, in this context. Semantic information provides a similarity relation between the meaning of words, –like synonymy. It is used here in order to provide AlvisRE with more ways of deciding on the similarity or dissimilarity of two candidate relations, or paths.

Different types of semantic information have been considered and tested with AlvisRE.

- Manually produced non-hierarchical semantic classes –simply called "manual semantic classes" forward.
- Traditional sources like WordNet, which while being manually produced have a sophisticated hierarchy and
- Distributional semantics systems, such as DISCO [Kolb, 2009] and word2vec [Mikolov et al., 2013a, Mikolov et al., 2013c].

Ontologies have not been used with AlvisRE, but the approach would be similar to that used for WordNet.

### 3.5.2 Manual Classes

The way the term semantic classes will be used in the context of this work corresponds to groupings of words that have similar meanings in the context of a specific RE task. These manually produced groupings are very limited in size but can serve as a proof of concept for the use of semantic information, as well as a baseline for comparison and evaluation of automated tools. In smaller corpora with limited variation in expression, manual semantic classes can be a reasonable approach. A simple similarity function is defined on semantic classes, which returns a score of 1 if two words belong to the same semantic class, and 0 otherwise.

### 3.5.3 WordNet

WordNet is a large lexical resource for the English language [Fellbaum, 1998]. WordNet groups lexical items in synonym sets, called synsets. It contains nouns, verbs, adjectives and adverbs. Synsets share semantic and lexical relations. The links between the synsets express conceptual-semantic and lexical relations. Examples of some of these hierarchical relations are hyponymy/hyperonymy, and meronymy.

For the calculation of semantic similarity based on distance between items in the WordNet lexical database AlvisRE uses the WS4J (WordNet Similarity for Java) toolkit, which is based on a famous similar toolkit for Perl [Pedersen et al., 2004]. Similarity between two items of WordNet is generally calculated by a topological similarity function measuring the length of the path connecting the two items.

WS4J includes the following algorithms:

- HSO [Hirst and St-Onge, 1998] which measures similarity based on the paths between the synsets of each word, provided that these paths are of reasonable length.
- LCH [Leacock and Chodorow, 1998] measures similarity as the shortest path between synsets by only taking into account WordNet's "IS-A" links and normalizing by scaling the path relatively to the depth of the taxonomy.
- LESK [Banerjee and Pedersen, 2002] is an adaptation to WordNet of the Lesk (1985) idea that the similarity between words can be calculated based on the overlaps of their definition.
- WUP [Wu and Palmer, 1994] takes into account the depths of the synsets of the two words as well as the depth of their lowest common subsumer.
- RES [Resnik, 1995] uses the information content of the lowest super-ordinate (most specific common ancestor) of the two words.
- JCN [Jiang and Conrath, 1997] computes a semantic distance combining the edge counts of "IS-A" links and the information content values of the concepts.
- LIN [Lin, 1998] is similar to JCN but uses a revised equation.

- Finally PATH uses the simple shortest path between words in the WordNet graph.

All the algorithms supported by the WS4J toolkit were tested, and HSO was chosen as the default choice.

WordNet focuses on general English and is not adapted to biomedical terminology. This means that often words that are frequent or meaningful in biomedical text are not found in WordNet or even when they are, their meaning in the context of biomedical literature is particular and not necessarily well-represented in WordNet. However, WordNet remains a very important lexical tool for Natural Language Processing applications and it is included in this study in order to confirm and show how a non-specific but widely recognized resource can influence the results.

### 3.5.4 Textual similarity

When considering using semantic information, I was faced with the problem of the limited scope of curated collections such as WordNet. Even tools such as DISCO frequently could not find the words I was searching for. By using textual similarity I hoped to fill this gap, based on the fact that words sharing etymology and meaning would have a higher textual similarity score than completely unrelated words, especially if some normalization like lemmatization, stemming or canonicalization takes place before the comparison.

For this I used textual similarity methods provided by the Java library Second String [Cohen et al., 2003]. Namely I used the Levenshtein distance, the Jaro or Jaro-Winkler distance [Winkler, 1990, Jaro, 1989, Jaro, 1995], the Smith-Waterman distance [Smith and Waterman, 1981] and the Needleman Wunsch distance [Needleman and Wunsch, 1970]. Additionally I used Dice's coefficient [Dice, 1945, Sørensen, 1948]. All of these similarity methods are variations on string edit distance. I tested these measures both on canonical and surface forms (see section 3.6.2).

### 3.5.5 Distributional Semantics

#### Introduction

The Distributional Hypothesis states that words with similar distributional properties have similar meanings [Harris, 1954]. Similar distributional properties mean that these words occur in the same contexts. Distributional Semantics (DS) is the domain of study which seeks to use statistical patterns of use of words by humans in order to understand their meaning. Other names for DS include Statistical Semantics, Corpus-based semantics, Geometrical models of meaning and Vector Semantics.

Distributional semantics have become a very popular method and it offers the possibility of learning from domain-specific corpora. This can guarantee that the words for which a RE system seeks semantic information will be proportionally represented in the used semantic source. Additionally, they solve the problem of a difference of meaning between a term used in everyday language and the biological domain. However, as these tools are purely automated and have no manual curation phase, they can contain errors linked to sampling biases or pre-processing. Finally, although traditionally using words, distributional semantics systems can also work with multi-word terms.

Two different approaches to DS have been tested with AlvisRE: DISCO, a traditional DS tool, and word embeddings, which are vectorial representations of words calculated from large corpora by using neural learning methods.

### DISCO

DISCO (extracting DIstributionally related words using CO-occurrences) is a tool that calculates a similarity score between words [Kolb, 2008, Kolb, 2009]. As the name implies, DISCO uses the co-occurrence of words as a way to identify their semantic similarity. DISCO cannot be trained on a user-specified corpus[23], but various versions are provided that have been trained on different corpora and on different languages.

Two of the possible versions have been selected. The first one was trained on a corpus from English Wikipedia and the second one on a corpus consisting of 100,000 articles of PubMed Central, which includes scientific articles from the biomedical database MEDLINE, as well as other life science journals literature. The Wikipedia corpus is of a much more important volume and has the advantage of additional pre-processing. Despite the smaller size, the PubMed one can promise better results based on the fact that it includes scientific articles, similar in nature to the source text of the corpora that we use to test, as well as the fact that it is specific to the biomedical domain.

### DISCO Wikipedia

The "Wikipedia" data pack is based an a version of English Wikipedia from April 2013. It contains approximately 1.9 billion tokens and 420,184 queryable words and multiword lexemes like "take_off", american_national_biography" or "forest_lawn_memorial_park". The corpus has been tokenized, stop words and words with a low frequency (<50) have been removed and the rest were converted to lowercase. The multi-word lexemes (mwl) have been identified using the SPECIALIST Lexicon, a large syntactic lexicon of biomedical and general English and spaces between words have been replaced by an underscore character. Less frequent multiword lexemes (frequency <50) have been equally removed, and a part-of-speech filtering has been applied, leaving only those that match a phrasal verb or a noun phrase pattern. In order to comply with this corpus format, AlvisRE transforms multi-word canonical forms by replacing spaces by underscores.

### DISCO Pubmed

The "PubMed" data pack is based on approximately 100,000 articles from the PubMed Open Access database taken in July 2007. It is significantly smaller than the Wikipedia one, with a total number of 181 million tokens and it includes 60,000 words that can be queried. In terms of pre-processing, the corpus has been tokenized and frequent words of small relative significance have been removed. This corpus is not lemmatized, so AlvisRE uses the surface forms of the words and not the lemmas for semantic lookup.

[23] The authors of DISCO were contacted at the time regarding this but did not reply. In 2016 a trainable version of DISCO was made public, but it was not used for this thesis.

### Word Embeddings

Word embeddings are vector space representations of words, which capture the word meanings. They are computed using neural network approaches. They were first introduced by [Bengio et al., 2001, Bengio et al., 2003] and recently regained the interest of the NLP community with the publication of word2vec [Mikolov et al., 2013a, Mikolov et al., 2013b].

The word2vec toolkit contains two log-linear neural network models for computing word embeddings, both based on words and their neighbours: the CBOW and the skip-gram models. Given a word w and the words around it in a window of size n, the CBOW model predicts the word w, given its neighbours in the window. The skip-gram model, inversely, predicts the neighbouring n words of the window, given the current word w. Both models can be trained with or without negative sampling.

Shallow neural learners were chosen based on the idea that trading the model complexity for efficiency, one can learn from much bigger datasets. The skip-gram model with negative sampling [Goldberg and Levy, 2014] is recommended by word2vec creators as the default choice. As word embeddings are vectors in the Euclidean space, the cosine similarity function can be applied to them.

As with DISCO, I produced two different versions, one based on Wikipedia and one on PubMed Abstracts. However, these corpora were not the same as the ones used for the training of DISCO, as those are older instances that are currently unavailable for download.

#### word2vec wikipedia

The texts from this source are wikipedia articles, packed in a dataset [Mahoney, 2011]. They correspond to the “first billion characters from wikipedia” and they were extracted from Wikipedia in 2006. This set is 1GB in size and contains ~123 million words, corresponding to a vocabulary of ~220 thousand words. This dataset is often used with word2vec as a source for general domain text.

#### word2vec pubmed

I created this dataset with the help of Robert Bossy. It is based on an instance of the database of Pubmed Abstracts from 2014. It contains approximately 20 million abstracts, totalling 15G in size. This dataset contains ~2 billion words and vocabulary of ~1 million unique words.

Since no lemmatization was done on these corpora, AlvisRE uses the surface form of words. No multi-word terms were used as these terms require additional preprocessing of the large word2vec corpora.

### Combining Sources

Ontologies and lexica such as WordNet contain lemmatized items. When working with large corpora such as those used for Distributional Semantics, it is very expensive to add linguistic processing steps, such as lemmatization or fusion of multi-word terms. Manually constructed semantic classes, even though they are of smaller size, are in practice much easier to produce when using lemmas. Any of these resources only provides a limited coverage of the full potential of human expression.

AlvisRE can combine multiple semantic similarity measures. It retains the maximum score. By combining multiple sources it can maximize the probability of finding a pertinent similarity value for a pair of words. In all of the above semantic tools, while a “good” similarity between two words means that at least in some context, these words are proven to have similar meanings, a “bad” similarity can be simply the result of one or both of them being underrepresented in the source. Consequently, choosing to keep the best similarity is a safer choice than using an averaging method.

## 3.6 Results

### 3.6.1 Experimental setup

#### Alvis Suite

Chapter 1 includes an overview of all the tools available in the Alvis Ecosystem (1.7). This section will briefly re-introduce the tools used for these experiments, along with any precisions and parameters that are relevant.

The Alvis Suite is a generic text-mining pipeline based on linguistic and machine learning technologies, that can be easily configured for specific domain applications. It includes, among other modules, AlvisNLP/ML [Nédellec et al., 2009, Ba and Bossy, 2016], a workflow for the biology domain, and AlvisRE. These two modules together perform a full semantic annotation of structured entities and n-ary relations in texts in the biology domain.

Alvis NLP/ML is a generic pipeline for the semantic annotation of textual documents. It integrates Natural Language Processing (NLP) tools for sentence and word segmentation, named-entity recognition, term analysis, semantic typing and relation extraction. These tools rely on resources such as terminologies or ontologies for the adaptation to the application domain. Alvis NLP/ML contains several tools for (semi)-automatic acquisition of these resources, using Machine Learning (ML) techniques.

High-quality prediction of entities is achieved by named-entity recognition and terminological analysis. Tools developed specifically in this spirit and integrated in AlvisNLP/ML include the term extractor BioYaTeA [Golik et al., 2013, Golik et al., 2011], ToMap and OnToMap for term categorization [Bossy et al., 2015, Golik et al., 2012b, Ratkovic et al., 2012, Nédellec et al., 2014]. Syntactic parsing is performed by a number of integrated parsers, such as CCG [Clark and Curran, 2007] and AlvisGrammar [Ratkovic, 2014, Ratkovic et al., 2012]. Coreference resolution is possible by creating dedicated AlvisNLP/ML scripts, such as the ones done for Bacteria Biotopes [Ratkovic, 2014, Ratkovic et al., 2012]. Other AlvisNLP/ML modules include textual segmentation, POS tagging and all other necessary steps for IE.

### 3.6.2 Experimental validation

During the development of AlvisRE all experimentation used real data from challenges, in particular the LLL challenge [Nédellec, 2005b] and the BioNLP Bacteria Biotope challenge.

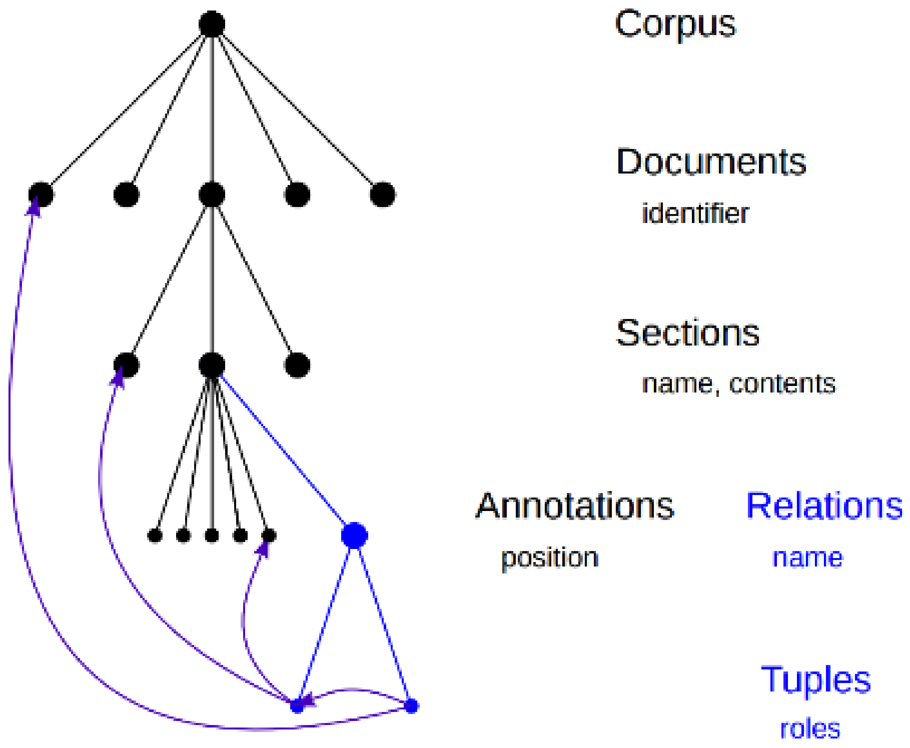


Figure 3.14: The data model of AlvisNLP/ML. AlvisNLP/ML uses a data model very similar to that of AlvisRE. Sections in AlvisNLP/ML correspond to sentences in AlvisRE, and Tuples correspond to relation arguments. Both types of relations possible in AlvisRE (syntactic and semantic) are modeled as Relations in AlvisNLP/ML. Annotations englobe all textbound types of AlvisRE, namely words, named entities, POS tags and other layers of textbound information.

These datasets are introduced in the second section of Chapter II, where various IE corpora are discussed. In order to provide the necessary context for the results, they will be briefly described in this section equally. Given that the data come from past challenges, they are particularly useful for comparison to the state of the art.

The testing of various scenarios and parameter optimization was also done on these corpora. This allowed me to evaluate the usefulness and performance of the various ideas explored during this project, and to provide adequate default values to the different possible parameters. In this section, selected scenarios and optimizations will be presented, aiming to explore and explain some of the most important aspects of AlvisRE.

This section does not cover the results on the newly produced *A. thaliana* corpus, which will be presented in a separate chapter. However, these intermediary experiments served as a guide for the choice of parameters used for the results on the *A. thaliana* corpus.

**A fix for missing dependency links.**

Experiments showed that dependency parsers would yield an important number of disconnected candidates. On the LLL dataset, the CCG parser was unable to produce complete trees for 20% of the candidate pairs: only 750 of the 926 possible pairs of candidates of the train set were connectable in the dependency graph. Preliminary test with other parsers such as the the Enju Parser gave even bigger disconnection rates.

In order to further understand the impact of the number of candidates on the prediction quality performance of AlvisRE, I set up a probabilistic study on the LLL corpus, where only a random percentage of the total number of all possible candidates of the entire dataset was used for cross-validation. The goal of this study was to estimate the loss of information due to the disconnected candidates. Since such disconnections in dependency parsings do not appear to follow some rules or regularity, I chose to emulate this phenomenon by keeping only a random number of all possible candidates of the entire dataset for cross-validation. To counter this probabilistic nature, cross-validation was repeated 20 times for percentages 10-90% (320 times in total) and another 320 times for percentages 90-100%, as this last window was found to have a much greater variance.

Figure 3.15 shows the results of this experiment. An observation that stands out immediately is the jump in all measures of about 50 points happening around 1250 candidates, or 83% of the whole set. Up to this point the increase in F-measure is steady at about 3 points for every additional 10% of candidates. Afterwards, results vary greatly until they stabilize at around 1400 (93%) where they achieve a score of 65, 70 and 60 for F-Measure, Recall and Precision respectively. These results confirm that the loss of candidates can have a very significant impact on the results. Moreover, in this setup 17% seemed to be the maximum loss that could allow for a relatively low negative impact, but as we saw earlier, even the best parser tested was unable to connect about 20% of the possible candidates. In the unfortunate event that this loss is high enough to fall under the "jump" threshold, the cost on the performance of the algorithm is dramatic. A solution to this problem seems crucial.

Once the solution of adding wordpath relations was implemented, tests confirmed that this approach does indeed improve results. For example, on the BioNLP '13 BB Task 2 dataset there was an overall ~2% gain in F-Measure using wordpath (see Table 3.1). Such improvements

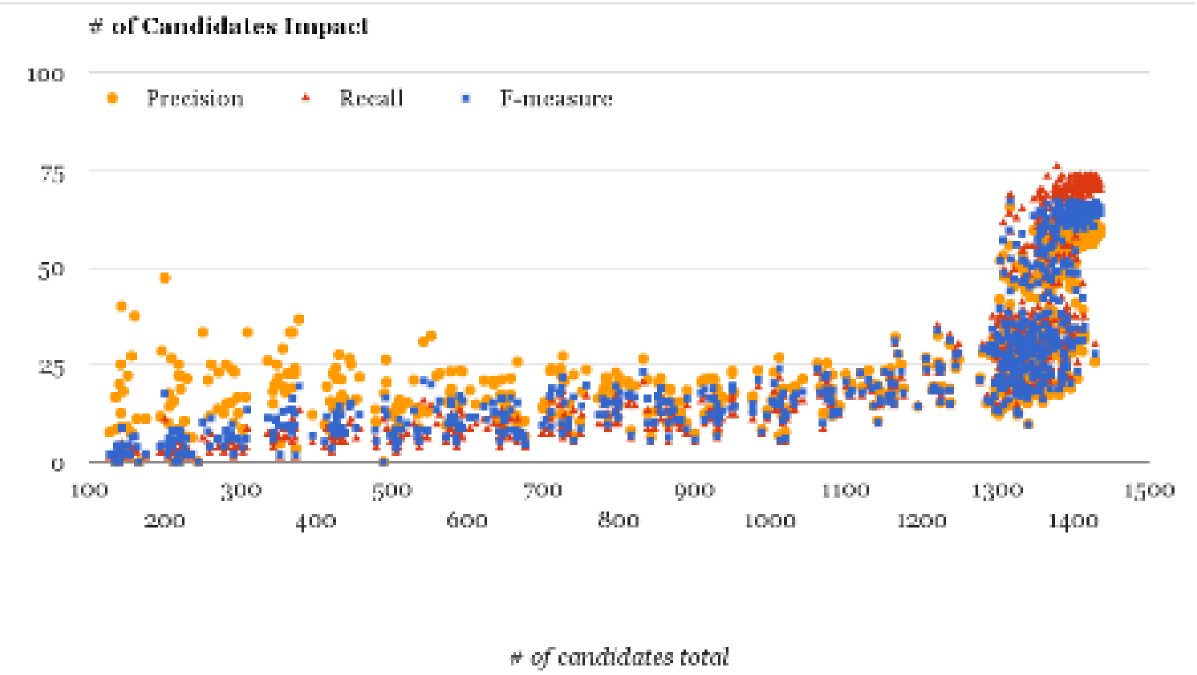


Figure 3.15: Study of the impact of the number of candidates on the performance of AlvisRE on the LLL corpus.

| Configuration | Precision | Recall | F-measure |
|---|---|---|---|
| Without wordpath | **52.7** | 63.1 | 57.5 |
| With wordpath | 51.4 | **70.0** | **59.3** |

Table 3.1: Results on the Bactetria Biotopes dataset using the challenge evaluation tool. This experiment used the native CCG parser and coreference resolution by Zorana Ratkovic.

are understandable once one looks at the number of disconnected candidates: 5,343 connected candidates for 6,278 possible couples (~85%) for the train+dev sets and 2,123 out of 2,650 possible (~80%) for the test set. Specifically for the test set, this amounts to ~15 positive relations that the algorithm would be incapable to predict, as their arguments were not connectable in the dependency graph. These results were produced using coreference resolution provided by Zorana Ratkovic [Ratkovic et al., 2012, Ratkovic, 2014] and optimized for the various parameters available (like the classifier C value). The official challenge tool [24] was used for evaluation.

As we can see in Table 3.1, adding wordpath slightly lowers overall precision. An intuitive explanation for this phenomenon is that by adding wordpath the classifier uses slightly heterogeneous representations, and thus its learning precision is penalized. On the other hand, adding wordpath links significantly improves the recall of the classifier, as it was expected. In fact, recall is increased enough for the loss of precision to be justified, as the combined measure of these two, the F-measure is indeed improved. However, in applications where the highest possible precision is necessary, this phenomenon should be taken into consideration before using wordpath links.

[24] http://genome.jouy.inra.fr/~rbossy/cgi-bin/bionlp-eval/BB_fix.cgi

| Representation | Co-occurrence | | | Co-occurrence+ Trigger Words | | | Global Alignment | | | Global Alignment+ Trigger Words | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PRE | REC | F-M | PRE | REC | F-M | PRE | REC | F-M | PRE | REC | F-M |
| EntireSentence | 12.6 | **98.7** | 22.4 | 12.9 | **98.7** | 22.8 | 12.8 | 74.6 | 21.9 | 16.5 | **79.5** | 27.4 |
| SurfacePath | 12.6 | **98.7** | 22.4 | **22.1** | 93.9 | **35.8** | 35.4 | **81.9** | 46.4 | 34.2 | 78.3 | 47.6 |
| DependencyPath | **14.4** | 93.9 | **25.0** | 18.7 | 92.7 | 31.2 | **58.1** | 77.1 | **66.3** | **60.1** | 78.3 | **68.0** |

Table 3.2: Results on the LLL corpus

### Representation and Similarity Function

Two series of experiments were conducted in order to test and compare different representation scenarios, as the latter were presented in section 3.3 of this chapter. Three path representations were used: word-order paths for the entire sentence (“EntireSentence” in the results tables), word-order paths between the candidate arguments (“SurfacePath” in the tables) and dependency paths using both dependency and wordpath relations (“DependencyPath” in the tables). Additionally, I considered the use of trigger words for these experiments. As the Global Alignment method allows for use of typical feature choices thanks to the Empirical Kernel Map transformation, testing with trigger words was done. Trigger words are very often used with co-occurrence in order to improve its precision, and this allowed me to compare Global Alignment with co-occurrence more fairly.

Tables 3.2 and 3.3 show the results of these tests on the LLL and BioNLP ’13 BB Task 2 corpora. These tests did not use wordpath links for the DependencyPath scenario, and no other parameter optimization was performed. Evaluation scores were computed by the corresponding challenge tools. The list of the trigger words used for each task are: for LLL the list of verbs constituted by Philippe Veber from the Bibliome team for the construction of manual semantic classes (Appendix A), and for BB: a list of trigger words constituted by Zorana Ratkovic and Robert Bossy for this task (Appendix B).

Table 3.2 shows the results on the LLL corpus. While normally a 100% recall is expected for the representations using co-occurrence on the sentence (EntireSentence and SurfacePath), a single example was not predicted correctly due to a simple bug[25] in the AlvisRE system at the time of these experiments. It explains why the recall is 98.7% and it does not affect the other measures. In spite of this bug, these results hold merit, as they still allow for comparison between the different representations.

In the first scenario, **Co-occurrence** requires that the two candidate arguments co-occur simply in the corresponding path representation, i.e. EntireSentence, SurfacePath or DependencyPath. Since no other constraint exists, this approach always gives the best recall, but lower precision, due to the fact that candidate arguments may co-occur for other reasons than being in an actual interaction relationship. The recall for the DependencyPath representation is lower than for the SurfacePath representation, due to the fact that disconnected sub-graphs of the dependency tree are not remedied by wordpath links. However, fewer examples seem to give better precision and consequently a better overall score for F-measure was achieved, but I do not believe there is more significance to this improvement.

**Co-occurrence with trigger words** requires the co-occurrence of the candidate arguments as well as at least one trigger word in the corresponding path representation. In the case of

[25] This bug was corrected in later versions of AlvisRE.

| Representation | Co-occurrence | | | Co-occurrence+ Trigger Words | | | Global Alignment | | | Global Alignment+ Trigger Words | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PRE | REC | F-M | PRE | REC | F-M | PRE | REC | F-M | PRE | REC | F-M |
| EntireSentence | 23.5 | **78.2** | 36.1 | **23.4** | **78.2** | **36.1** | 51.0 | **65.9** | **57.5** | 51.0 | **65.9** | **57.5** |
| SurfacePath | 26.5 | **78.2** | 36.1 | **23.4** | **78.2** | **36.1** | 50.4 | 36.8 | 56.3 | 50.3 | 63.8 | 56.2 |
| DependencyPath | **27.7** | 75.9 | **40.6** | 17.4 | 27.4 | 21.3 | **51.6** | 63.3 | 56.8 | **51.6** | 63.3 | 56.8 |

Table 3.3: Results on the BB Task 2 corpus (with anaphora resolution)

EntireSentence, this does not appear to influence recall, since the list of trigger words used appears to be permissive enough to not penalize any true positive example. However trigger words only slightly filter false positives, only enhancing the precision by little. Adding the requirement that these trigger words occur between the two candidate arguments in the sentence for the SurfacePath representation lowers the recall, missing some true positives, but at the same time improves precision enough for the overall performance (F-measure) to be improved significantly by 13 points from 22.4% to 35.8%. While the effect on the DependencyPath representation follows similar tendencies as it did for SurfacePath, simple co-occurrence on the DependencyPath appears to not be as good a predictor as it was for SurfacePath, yielding lower scores for both precision and recall compared to the latter.

The situation is quite different when the similarity function used is **Global Alignment instead of** co-occurrence. Aligning dependency paths seems to be the best approach overall as it is so much more precise that even with the loss of recall, the resulting F-measure scores are significantly superior (resp. 66.3% and 68%) in both Global Alignment scenarios (with and without trigger words) with the scores of DependencyPath being roughly triple of those for the EntireSentence representation. Between the two representations using the surface form of the sentence, keeping only the part of the path between the arguments (SurfacePath) doubles the F-measure, which makes it a valid alternative when computation time is a concern, thus making the use of a dependency parser costly. While this representation does not really compare to the DependencyPath results, which give an additional rough 20% of F-measure, it can still be considered as it is much faster than DependencyPath.

Finally, the **Global Alignment with trigger words** scenario is taking advantage of the possibility to integrate features in the Global Alignment approach by adding the existence of trigger words as such features. These extra features do not appear to have a generalized positive effect on the surface representations, but they increase both precision and recall for the DependencyPath representation. While this set of experiments is not large enough to be able to make definitive conclusions, it proves that features used traditionally in feature-based RE approaches can indeed be integrated in AlvisRE and in some cases improve results, especially in the case of dependency representations, which are the representations around which AlvisRE was built.

Table 3.3 shows the results of the same set of experiments done on the BioNLP '13 BB Task 2 corpus. Overall, the results show the same characteristics in most cases as for the LLL corpus. The differences for BB are the following: a) trigger words do not appear to make a difference for the SurfacePath representation when using co-occurrence, and greatly penalize both precision and recall for the DependencyPath representation, b) the loss of recall for the DependencyPath because of disconnected paths is such that even with slightly better precision, it does not give the best F-measure scores between the three when using Global Alignment, and c) trigger words appear to have no effect when using Global Alignment, probably due to the high dimensionality

| Participant | Precision | Recall | F-measure |
|---|---|---|---|
| TEES-2.1 | 0.82 | 0.28 | 0.42 |
| IRISA-TexMex | 0.46 | 0.36 | 0.40 |
| Boun | 0.38 | 0.21 | 0.27 |
| LIMSI | 0.19 | 0.04 | 0.06 |

Table 3.4: BioNLP'13 BB Task 2 participant evaluation.

of the transformed vectorial representation that makes a single feature carry proportionally less weight than in the case of LLL, where there are much fewer training examples and, thus, lower dimensionality.

By looking at the data in both corpora, I noticed a difference between the two sets. LLL is a smaller corpus and contains phrases that are much more homogeneous among them than the larger BB corpus. The conclusion to which I arrived is that the similarity of results regardless of different representations for BB using Global Alignment is due to the fact that there exists a cluster of relatively homogeneous sentences, both syntactically but also in their surface form, which all representations manage to capture. But apart from this cluster, the rest of sentences are so dissimilar that none of these representations manage to allow the Global Alignment algorithm to correctly classify them.

The general conclusion of these experiments for both corpora is that Global Alignment is clearly superior to co-occurrence methods. Regardless of the representation chosen, in both cases Global Alignment approaches gave 20-30% better in F-measure than the best co-occurrence result. It is worth noting that this method gives better results than the state of the art. Table 3.4 contains the official scores of the participants in Task 2 of the recent BioNLP '13 challenge [Bossy et al., 2015], where the best F-measure obtained was 42%, 15 points less than what was obtained in this set of experiments. Furthermore with the wordpath reparations and other parameter optimizations done, AlvisRE reaches an F-Measure of 59.3% (see table 3.1), putting it 17.3% ahead of the best participant score.

It is important to mention that according to the task organizers, most participants had not performed coreference resolution in their submissions to this task. Unfortunately, revised result tables for the participants using coreference resolution are not available at this time. While the focus of this section is not the comparison with the task participants, if the reader is interested in understanding the role of each part of our system in these scores, she or he is referred to the thesis of Zorana Ratkovic [Ratkovic, 2014], where she compared and examined the improvements of coreference resolution and wordpath for AlvisRE-Global Alignment (SPGAK in the referenced thesis) on this task and found that the former improved the F-measure by almost 37 points and the latter by 2 additional points.

### Semantic Information

In order to document my work in studying the use of semantic information in AlvisRE presented in section 3.5 of this chapter, two sets of experiments will be presented in this section: a) one testing WordNet, DISCO and textual similarity and b) a second one using manual semantic classes and word embeddings. I used the LLL corpus for both and came to the conclusion that using semantic similarities can indeed improve RE methods, as their use with AlvisRE on LLL proved to boost F-Measure up to 5 points by raising Recall and even Precision.

The first experiment represents some of the first steps in my integration and evaluation of semantic information. I evaluated various WordNet similarity measures, DISCO distributional semantics, and textual similarity as a substitute for semantic similarity. The configuration for this set was AlvisRE with the CCG parser and no wordpath links, and cross-validation on the entire corpus. The results can be found in Table 3.5.

For textual similarity, I tested the algorithms provided by the SecondString library (section 3.5.4) as well as Dice's coefficient and found that the best score was obtained by using the latter on the surface representation of the words. Most of these textual similarity measures lowered the precision while maintaining the recall roughly the same. Dice's coefficient was actually the measure that improved the recall enough to obtain an F-Measure score that was slightly improved. The way textual similarity can help AlvisRE is by giving a non-binary response to the similarity of two words, so words sharing some etymological roots get a better score than completely different ones. This translates in better recall. On the other hand, visually similar but unrelated words can have better scores than they would in a binary scenario, thus lowering the precision. Using lemmas and canonical forms amplified the increase in recall and decrease in precision. A possible explanation could be inconsistency in the form and frequency of such normalizations, but I did not investigate this phenomenon further.

While textual similarity cannot really be a substitute for semantic similarity, it is useful as often semantic similarity resources have a limited vocabulary. It was in this spirit that the combination of textual similarity and semantic similarity was tested. I chose to combine the best performing textual measure (Dice's coefficient) with WordNet, DISCO and a combination of both. For DISCO I used the Pubmed version (section 3.5.5), as previous tests had shown that a domain-specific resource gave better results.

Surprisingly, using WordNet consistently decreased both precision and recall, something I attribute to the fact that WordNet is not adapted to the domain. Even if some words in a general context are similar semantically, it is often the case that in the biomedical domain only some of their possible meanings are pertinent. While using DISCO only raised both precision and recall by approximately 1 point, it was an encouraging development in my effort to integrate and evaluate semantic similarity measures. The overall gain from the baseline was an F-Measure increase of 1.5 points.

The lessons learned from previous experiences not included in this thesis were that it was important to use domain-specific resources but also that these needed to be as complete as possible. Coincidentally, word embeddings started becoming popular during that period.

For the second experiment I chose to avoid using textual similarity altogether and compare two diametrically opposite semantic resources: manually crafted semantic classes and a large-scale, automated, distributional semantics tool: word2vec. The manual semantic classes were previously constructed by Philippe Veber, by grouping together 110 verbs into 22 classes; the extensive list can be found in Appendix A. While these classes are very limited in scope, they were created specifically for LLL. For the word2vec scenario, I trained word2vec on an input of around 20 million Medline abstracts (Medline 2014 version), ensuring a good vocabulary coverage and domain specificity. For evaluation, I used the official challenge tools.

In spite of their small size, manual classes increased precision by ~1 point and recall by ~4 points, bringing the F-measure up by ~2 points. This result highlights the importance of using

| **Similarities used** | Precision | Recall | F-measure |
|---|---:|---:|---:|
| *Baseline* | 66.6 | 60.2 | 63.2 |
| Jaro on surface | 52.6 | 60.2 | 56.1 |
| Levenstein on surface | 60.9 | 60.2 | 60.6 |
| Smith Waterman on surface | 53.6 | 61.4 | 57.3 |
| Needleman on surface | 65.7 | 60.2 | 62.8 |
| Needleman on canonical | 65.7 | 57.8 | 61.5 |
| Dice on surface | 61.7 | 66.2 | 63.9 |
| Dice on surface & canonical | 60.2 | 67.4 | 63.6 |
| Dice on canonical | 60.2 | 67.4 | 63.6 |
| Dice surface + Wordnet PATH | 57.2 | 66.2 | 61.4 |
| Dice surface + Wordnet LIN | 60.4 | 62.6 | 61.5 |
| Dice surface + Wordnet LESK | 59.7 | 62.6 | 61.1 |
| **Dice surface + Disco 1st order** | **62.0** | **65.0** | **63.5** |
| Dice surface + Disco 2nd order | 62.2 | 67.4 | 64.7 |
| Dice surface + Wordnet HSO | 61.3 | 65.0 | 63.1 |
| Dice surface + max (HSO. DISCO) | 60.8 | 67.4 | 64.0 |

Table 3.5: A study on LLL using AlvisRE with Global Alignment and semantic similarities.

| **Setup** | **Precision** | **Precision** | **F-measure** |
|---|---:|---:|---:|
| *baseline* | 56.4 | 73.4 | 63.8 |
| manual classes | 57.6 | 77.1 | 65.9 |
| word2vec medline 2014 | **63.9** | **74.6** | **68.8** |
| word2vec medline 2014 + wordpath | **63.3** | **77.1** | **69.5** |

Table 3.6: Using manual classes and word embeddings on LLL

adapted vocabulary and measures for each domain, as even with such a limited scope there is a significant performance improvement. However, the definitive winner of this comparison is word2vec. Inversely from the manual classes, recall was only a little improved (~1 point), while there was a 5 point increase in precision, resulting in a 5 point increase in F-measure. Additionally, Table 3.6 includes the scores obtained when adding wordpath links to this configuration, which added almost 1 point more.

As a final note on my experiments with semantic similarity, I would like to add that while similar experiments were performed for the BB corpus, the results were not improved by any type of semantic information. As a matter of fact, my results on BB seemed as impervious to semantic similarities as they were with regards to parameter optimization and trigger words. The scores were only affected by the choice of representation and similarity function for AlvisRE, as it was shown in section “Representation and Similarity Function”, where I also theorized about the reasons behind this behaviour.

### A complete IE task

For the last set of results presented in this section, that were obtained by AlvisRE on a complete IE task, I have chosen BioNLP’13 BB Task 3: a task for which the Named Entities were not

| Relations<br>in a sentence | Relaxed<br>biotope boundaries | Relaxed<br>bacteria boundaries | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| | | | 23.2 | 30.1 | 26.2 |
| ✓ | | | 16.0 | 27.3 | 20.1 |
| | ✓ | | 33.1 | 38.9 | 35.8 |
| | | ✓ | 35.7 | 39.0 | 37.3 |
| ✓ | ✓ | | 21.5 | 35.0 | 26.7 |
| | ✓ | ✓ | 49.4 | 50.3 | 49.8 |
| ✓ | | ✓ | 24.0 | 33.1 | 27.8 |
| ✓ | ✓ | ✓ | 32.4 | 42.7 | 36.8 |

Table 3.7: Results on BB Task 3. Entities prediction by Louise Deleger and the Bibliome team. AlvisRE was trained on train+dev and tested on test.

given, but which required a system capable of performing both NER and RE. While I personally only worked on relation extraction, I believe that it is important to evaluate a RE system when its most crucial input, the relation arguments, is not manually annotated, but itself the result of a prediction. This way, the RE system is tested in a scenario resembling a real-world application.

For the NER recognition, Alvis NLP tools were used: entity boundaries were identified using BioYatea [Golik et al., 2013, Golik et al., 2011], bacteria prediction was done using a dictionary of the NCBI taxonomy[26] [Federhen, 2012], and habitats were predicted by using a version of ToMap [Golik et al., 2011] adapted and extended by Louise Deleger. Lemmatization was done using the Genia tagger [Tsuruoka et al., 2005], and CCG was used for both POS tagging and dependency analysis. Anaphora resolution was done by Zorana Ratkovic, using the methods described in her thesis for the Bacteria Biotopes corpus [Ratkovic, 2014]. As far as the configuration of AlvisRE is concerned, default parameters chosen based on other tasks (LLL and BB Task 2) were used for Global Alignment on dependency paths enriched with wordpath links. No semantic similarities were used. The results presented here were obtained using the official challenge evaluation tool [27]. This tool has several evaluation options, and those relevant to RE will be included in the results. Namely,

- Only evaluate relations included in a sentence
- Do not count habitat boundaries in the evaluation
- Do not sanction wrong bacteria boundaries

By examining the results (Table 3.7) of AlvisRE on this task, we can draw the following conclusions. Firstly, concerning inter versus intra sentence relations: since the training was done using all relations and not only those that lie within one sentence, our results are better when testing on all relations as compared to testing on intra-sentence relations only. As coreference resolution greatly improved the scores AlvisRE obtained on the RE-only Task 2 of the same challenge 3.1, it was considered without merit to submit a prediction without coreference resolution. Second, relaxed boundaries on named entities naturally gave better scores. NER for this challenge was a difficult task, so reducing the penalization of imperfect predictions expectedly helped obtain better results. As a matter of fact, the best F-Measure (49.8%) is indeed obtained when testing on all relations with relaxed boundaries for both types of named entities.

[26] http://www.ncbi.nlm.nih.gov/taxonomy

[27] http://genome.jouy.inra.fr/~rbossy/cgi-bin/bionlp-eval/BB_fix.cgi

| Official Scores | | | | Relaxed biotope boundaries | | | Relaxed bacteria boundaries | | |
|---|---|---|---|---|---|---|---|---|---|
| **Participant** | **PRE** | **REC** | **F-M** | **PRE** | **REC** | **F-M** | **PRE** | **REC** | **F-M** |
| TEES-2.1 | 18 | 14 | 14 | 61 | **41** | **49** | 52 | 28 | 36 |
| LIMSI | 12 | 4 | 6 | **82** | 9 | 15 | **71** | 7 | 10 |
| AlvisRE | **23** | **30** | **26** | 33 | 39 | 36 | 36 | **39** | **37** |

Table 3.8: BioNLP'13 BB Task 3: The official scores of the participants in the challenge as they appear in the official overview publication, compared to AlvisRE.

When comparing with the official evaluation of the challenge [Bossy et al., 2015] (Table 3.8) our system obtains the best F-Measure score. In the relaxed entity boundary scenarios TEES 2.1 is superior (biotope) and equivalent (bacteria) to AlvisRE, obtaining a much higher precision and similar or inferior recall. The difference in scores between the normal evaluation and the relaxed boundary evaluations is much more important for the two participants than it is for AlvisRE. A possible explanation for this observation would be that our NER method had fewer boundary problems than those of the other participants. Official scores limiting testing to intra-sentence relations are not available for all relation types so they are not included in this table.

## 3.7 Discussion

This chapter outlines how AlvisRE takes input produced by previous IE steps, transforms its representation and uses it to learn and predict semantic relations. The chapter provides a number of intermediary results that guide the reader through the method components and the choices made while developing it.

AlvisRE is a modular method which can take advantage of different types of information, according to resource availability and constraints. It has been conceived and developed with dependency graph representations in mind and it achieves the highest performance using them. In order to avoid missing positive examples in both training and test due to dependency parsing errors, AlvisRE can make use of neighbourhood relations between words to link disconnected subgraphs of the dependency tree. However, when dependency parsings are not available or computational time is very limited, AlvisRE can use surface representations.

The classification method proposed by AlvisRE is based on an edit distance/global alignment approach. When used with dependency paths, the Global Alignment classifier achieves results that are state of the art. But even with surface representations, it has been shown to obtain results that are greatly improved over simplistic methods such as co-occurrence.

Using kernel methods has the privilege of avoiding the often counter-intuitive process of producing features that encapsulate all the desired knowledge. The Global Alignment method is a similarity function and thus a kernel. Nevertheless, it does not satisfy the mathematical properties necessary to be used safely directly with a kernel method such as SVMs. A transformation called Empirical Kernel Map is used to produce vectorial representations using the Global Alignment function. These representations are then safe to use with a kernel method and any traditional kernel (e.g. dot product). This transformation has the additional advantage of producing vectors which can be augmented by features, such as the occurrence of trigger words. Using trigger words improved scores on the LLL corpus, but this was not the case for the larger BB corpus. A possible interpretation is that the larger size of the BB corpus training set results in highly dimensional vectors and consequently reduces the relative weight of each feature.

The use of semantic information was evaluated on both LLL and BB. WordNet was found to be too limited in its scope and too generic in its purpose, as the semantic distances based on it had a negative impact on RE. Manually constructed semantic classes and distributional semantics, on the other hand, had a positive impact. Manual classes were very limited in scope but improved F-Measure by a little over 2 points. DISCO improved scores only slightly and was found to frequently not contain the terms looked up. Finally, the more recent word2vec tool trained on Pubmed abstracts gave a significant boost of 5 points to the F-measure. While all of these three semantic sources improved results on LLL, the results on the BB corpus were unaffected by them. The BB corpus appeared to be quite resistant to parameter optimization in general, leading me to believe there exists a cohesive, relatively homogeneous group of examples within it that responds well to the classification in any case, while the rest remains heterogeneous and difficult to correctly classify by the approaches tested. This corpus has the particularity of being based on wiki pages on biology, which might follow different standards of expression than scientific articles, but whatever the cause this solidifies the theory that there exists no free lunch in machine learning, and it calls for rigorous examination of the text and data used in order to adapt the methods applied. Experiments on the most recent BB corpus, BB'16, which contains Pubmed abstracts, could contribute to validate this hypothesis.

AlvisRE was also used in a scenario where the named entities were not given but were predicted by methods set up by my colleagues at INRA. The results of AlvisRE on the task of relation extraction were once again superior to those produced by the participants in this 2013 challenge, leading me to believe that AlvisRE is indeed competitive.

## 3.8 Conclusion

AlvisRE allows for a lot of modularity and integration of different approaches, which calls for hands-on experimentation and validation of all these possibilities. The tests I ran throughout my work on this topic guided me to the following conclusions:

1. In most cases, using dependency paths and Global Alignment is the best approach.
2. When computational performance is an issue, Global Alignment can also be used on the text of the sentence. While this configuration often produces results inferior to the ones obtained with dependency parsing, it can scale well and be more appropriate for some applications, while being significantly better than co-occurrence methods.
3. Adding wordpath links is a good solution to disconnected paths when using dependency-based representations.
4. Semantic information can significantly improve the results, but this should be decided on a per-corpus basis. Additionally, use of semantic information should be decided on a per-application basis since external resource lookups add to the computation time.

5. AlvisRE has been tested in real world scenarios where entity annotation and coreference resolution were not performed manually but were also predicted.
6. In all the corpora tested, AlvisRE outperformed the official results of competitions even without using all possible optimizations each time.

# Chapter 4

# Relation Extraction on *A. thaliana*

**4.1 Introduction . . . 119**

**4.2 SeeDev in BioNLP-ST '16 . . . 120**

**4.3 AlvisRE on *Arabidopsis thaliana* . . . 122**

4.3.1 Optimizing the SVM margin parameter . . . 122

4.3.2 Exploring different relation types . . . 122

4.3.3 The impact of model transformations on Relation Extraction . . . 124

4.3.4 AlvisRE compared to the SeeDev participants . . . 125

**4.4 Discussion . . . 126**

## 4.1 Introduction

The two axes of this work were the corpus on seed development in *Arabidopsis thaliana* and the Relation Extraction (RE) method AlvisRE. The motivation behind the creation of this corpus was discussed in chapter I and the corpus itself, along with its transformation were presented in detail in chapter III. While intermediary results used to develop AlvisRE were covered in chapter IV, this chapter will present and discuss the results of AlvisRE on the newly constructed corpus on seed development in *Arabidopsis thaliana.*

In particular, the results presented here are the ones obtained on the SeeDev corpus, the version of the corpus presented in chapter III which was used in the recent BioNLP 2016 Shared Task. The SeeDev challenge corpus is an ideal candidate for evaluation as it is the best curated edition of the *A. thaliana* seed development corpus, thanks to the efforts of the SeeDev task organizing team. It also has the privilege of having been used on an already published and evaluated challenge, giving me the possibility to compare my results with those of the other participants.

## 4.2 SeeDev in BioNLP-ST '16

The SeeDev task had two separate subtasks “SeeDev-binary” and “SeeDev-full” [28], corresponding to binary relation extraction and n-ary event extraction respectively. Only “SeeDev-binary” is covered by this thesis. For this task, participants were asked to predict relations between entities given as input. The evaluation of submissions was done by Recall, Precision and F-measure, and these statistics were calculated for all relations but also for each type separately, as well as for each category of relations -without taking into account the relation types, in order to best asses the strengths of each submission. Tables 4.1 and 4.2 list the official results for the challenge, in summary and in detail, respectively.

The participating teams were DUTIR (Dalian University of Technology, China), LIMSI (CNRS, France), LitWay (Xidian University, China), ULisboa (LaSIGE, Universidade de Lisboa, Portugal), UniMelb (University of Melbourne, Australia), VERSE (University of British Columbia, Canada) and UTS (University of Turku, Finland).

All the participants used supervised Machine Learning methods, with five of them using SVM-based approaches, one MaxEnt (LIMSI) and one convolutional neural networks (DUTIR). The best overall F-measure (0.432) was achieved by LitWay from Xidian University. LitWay employed a system which used SVMs for the prediction of some relation types, and hand-crafted rules for others.

While all teams used standard linguistic information, deep syntactic parsing was used by four teams and notably the three best ranked. Additionally, LitWay used word-embedding to build features for its SVM-based method. Among the participants, only the LIMSI system included co-reference resolution, a step necessary for predicting inter-sentence relations. However, relations spanning multiple sentences were rare enough (4%) in the test dataset that they had little effect on the quality of prediction. The results of all the teams were equivalent when the evaluation was restricted to relations occurring within a single sentence.

Different relation types appeared to behave differently for the participating teams. In addition to the number of examples available in the corpus, which always has an impact on machine learning techniques,we believe there were two aspects contributing to this phenomenon: a) the regularity of the expressions used for each type and b) the argument constraints defined by the relation signatures [Chaix et al., 2016c].

While we conducted no study on the heterogeneity of expression of each relation type in the corpus, the cardinality of type signatures was considered. Types with more restrictive signatures, such as "Composes Primary Structure", which has 8 (2x4) possible argument combinations, had only few examples (51) in the dataset. Yet, the highest F-measure for this type was 0.67, proving to be easier to predict than the other relation types. On the other extreme, occurrences of "Regulates Expression" were more numerous (450) but this type has a larger number of possible argument combinations (4x16), and proved challenging for the participants, with a highest F-measure of 0.39.

[28] This section summarizes the recently published overview of the task in the BioNLP '16 Workshop Proceedings [Chaix et al., 2016a]

| Participant | F-measure | Precision | Recall |
|---|---:|---:|---:|
| LitWay | **0.432** | 0.417 | 0.448 |
| UniMelb | 0.364 | 0.345 | 0.386 |
| VERSE | 0.342 | 0.273 | **0.458** |
| UTS | 0.335 | **0.533** | 0.245 |
| ULISBOA | 0.306 | 0.379 | 0.256 |
| LIMSI | 0.255 | 0.212 | 0.318 |

Table 4.1: The official results by the participating teams for SeeDev-binary [Chaix et al., 2016a], by participating team.

| Relation Type | Best<br>F-measure | Participant |
|---|---:|---|
| **All Relations** | **0.432** | **LitWay** |
| **Time and Localization** | **0.142** | **LitWay** |
| Exists At Stage | 0.167 | ULISBOA |
| Exists In Genotype | 0.492 | LitWay |
| Occurs During | 0.000 | - |
| Occurs In Genotype | 0.167 | VERSE |
| Is Localized In | 0.450 | LitWay |
| **Function** | **0.255** | **ULISBOA** |
| Is Involved In Process | 0.000 | - |
| Transcribes Or Translates To | 0.343 | VERSE |
| Is Functionally Equivalent To | 0.708 | LitWay |
| **Regulation** | **0.416** | **LitWay** |
| Regulates Accumulation | 0.316 | UniMelb |
| Regulates Development Phase | 0.376 | UniMelb |
| Regulates Expression | 0.386 | UniMelb |
| Regulates Molecule Activity | 0.000 | - |
| Regulates Process | 0.504 | LitWay |
| Regulates Tissue Development | 0.000 | - |
| **Composition and Membership** | **0.490** | **LitWay** |
| Composes Primary Structure | 0.667 | LIMSI |
| Composes Protein Complex | 0.500 | UTS |
| Has Sequence Identical To | 0.867 | LitWay |
| Is Member Of Family | 0.534 | LitWay |
| Is Protein Domain Of | 0.438 | LitWay |
| **Interaction** | **0.303** | **UniMelb** |
| Interacts With | 0.286 | UniMelb |
| Binds To | 0.310 | VERSE |
| **Specific to Binary Framework** | **0.154** | **VERSE** |
| Is Linked To | 0.154 | VERSE |

Table 4.2: The official results by the participating teams for SeeDev-binary [Chaix et al., 2016a], by relation type.

## 4.3 AlvisRE on *Arabidopsis thaliana*

AlvisNLP/ML was used to prepare the input for AlvisRE, as it was described in chapter 3. The selected configuration used AlvisGrammar/CCG for syntax and wordpath links. The parameters used for the configuration of AlvisRE were not optimized for this corpus, but I used the ones that had been chosen on previous corpora. Namely, parameters which were not optimized include the various weights which concern AlvisRE's *ppscore* function.

### 4.3.1 Optimizing the SVM margin parameter

Early tests with the *A. thaliana* corpus showed that training the classifier with each class/relation type separately was significantly better to using the "1-vs-all" multiclass classification which is standard for Support Vector Machines (SVMs). While experimenting with this corpus during training, I realized that the optimal C parameter[29] of the SVM –the soft margin– differed for each relation type.

I used 10-fold cross-validation on the train and development tests in order to study the effect of the C parameter for each relation type. I chose 10 folds following the advice of Hastie et al. [Hastie et al., 2001b], who recommend 5- and 10-fold cross validation for model selection. Table 4.3 contains the measures for the best C parameter for each class. For the rest of the options and parameters available in AlvisRE, the values and weights used were those that were proven to be best based on the results on BB and LLL (see section 3.6).

I performed this Cross-Validation experiment with two different setups: one on the whole seedev-binary dataset (train+dev+test) (Table 4.3) and one on the train+dev datasets. While for most entity types the optimal margin parameter was identical or within one order of magnitude, "Interacts With" had the most significant change. The optimal value for this type was 1 on the train+dev dataset, but 100 on the whole dataset. This difference in value means that the latter needed a tighter margin to avoid misclassification. This translates into a larger heterogeneity of positive examples (candidate "Interacts with" relations which were true), or a bigger similarity between positive and negative examples. These two scenarios are not mutually exclusive, and they are both possible results of adding more documents to the dataset.

### 4.3.2 Exploring different relation types

Given that "Interacts With" is a "cacth-all" type for any interaction that does not satisfy the criteria of more specific types [30], I believe it is one of types that are most susceptible to noise from negative examples. Participants in the challenge appear to have struggled with this relation type (see table 4.2), along with "Is Linked To", a type defined for the necessities of the binary transformation of the SeeDev corpus, making it another example of "catch-all".

Wishing to study the differences between the train,dev, and test datasets, I compared the results obtained by Cross-Validation on them. Table 4.4 lists the variation of results on some

[29] This parameter controls the penalization of misclassification of training examples. Larger values lead to tighter margins which minimize training misclassification, while smaller ones lead to larger-margin separating hyperplanes even if it means misclassifying some training examples.

[30] see Section 2.3.1.0

| **Relation Type** | **C** | **F-Measure** | **Precision** | **Recall** |
|---|---|---|---|---|
| **All Relations** | | **0.408** | **0.323** | **0.681** |
| **Time and Localization** | | **0.559** | **0.521** | **0.621** |
| Exists At Stage | 1.0 | 0.471 | 0.390 | 0.593 |
| Exists In Genotype | 10.0 | 0.597 | 0.577 | 0.618 |
| Occurs During | 1.0 | 0.692 | 0.643 | 0.750 |
| Occurs In Genotype | 0.1 | 0.190 | 0.116 | 0.517 |
| Is Localized In | 10.0 | 0.569 | 0.515 | 0.635 |
| **Function** | | **0.345** | **0.306** | **0.485** |
| Is Involved In Process | 100.0 | 0.708 | 0.767 | 0.657 |
| Transcribes Or Translates To | 0.1 | 0.264 | 0.161 | 0.735 |
| Is Functionally Equivalent To | 1.0 | 0.239 | 0.188 | 0.329 |
| **Regulation** | | **0.397** | **0.273** | **0.830** |
| Regulates Accumulation | 10.0 | 0.534 | 0.524 | 0.544 |
| Regulates Development Phase | 1.0 | 0.305 | 0.198 | 0.662 |
| Regulates Expression | 1.0 | 0.363 | 0.238 | 0.765 |
| Regulates Molecule Activity | 0.1 | 0.577 | 0.536 | 0.625 |
| Regulates Process | 0.1 | 0.413 | 0.264 | 0.948 |
| Regulates Tissue Development | 1.0 | 0.667 | 0.824 | 0.560 |
| **Composition and Membership** | | **0.285** | **0.192** | **0.562** |
| Composes Primary Structure | 0.1 | 0.316 | 0.204 | 0.708 |
| Composes Protein Complex | 1.0 | 0.491 | 0.382 | 0.684 |
| Has Sequence Identical To | - | 0.000 | 0.000 | 0.000 |
| Is Member Of Family | 1.0 | 0.369 | 0.241 | 0.780 |
| Is Protein Domain Of | 1.0 | 0.390 | 0.275 | 0.667 |
| **Interaction** | | **0.408** | **0.393** | **0.425** |
| Interacts With | 100.0 | 0.404 | 0.392 | 0.417 |
| Binds To | 1.0 | 0.414 | 0.394 | 0.436 |
| **Specific to Binary Framework** | | **0.507** | **0.507** | **0.507** |
| Is Linked To | 100.0 | 0.507 | 0.507 | 0.507 |

Table 4.3: Results obtained by AlvisRE using Cross-Validation on a the combined "train+dev+test" seedev-binary dataset.
*Note: AlvisRE did not produce results for “Has Sequence Identical To” due to technical problems.*

| Type | Best F-measure on train+dev | Best F-measure on train+dev+test | Improvement |
|---|---|---|---|
| Is Protein Domain Of | 0.229 | 0.390 | 41.28% |
| Exists In Genotype | 0.435 | 0.597 | 27.14% |
| Is Involved In Process | 0.524 | 0.708 | 25.99% |
| Occurs During | 0.526 | 0.692 | 23.99% |
| Binds To | 0.336 | 0.414 | 18.84% |
| Composes Protein Complex | 0.466 | 0.491 | 5.09% |
| Is Linked To | 0.492 | 0.507 | 2.96% |
| Regulates Accumulation | 0.539 | 0.534 | -0.94% |
| Is Localized In | 0.609 | 0.569 | -7.03% |
| Interacts With | 0.467 | 0.404 | -15.59% |
| Is Member Of Family | 0.437 | 0.369 | -18.43% |
| Composes Primary Structure | 0.438 | 0.316 | -38.61% |
| Exists At Stage | 0.696 | 0.471 | -47.77% |
| Is Functionally Equivalent To | 0.545 | 0.239 | -128.03% |
| Occurs In Genotype | 0.528 | 0.190 | -177.89% |

Table 4.4: The F-measure calculated by Cross-Validation on the train+dev, and train+dev+test datasets. The last column reflects the difference in F-measure between the former and the latter, as a percentage of the F-measure on the whole (train+dev+test) dataset. Positive numbers correspond to better results on the whole dataset than on train+dev.

representative relation types between the training datasets (train+dev) and the whole dataset (train+dev+set).

Two relation types saw an important decrease in performance, "Is Functionally Equivalent To" and "Occurs in Genotype", meaning that the test dataset introduced examples that were particularly difficult for AlvisRE to classify. The change on the rest of the relations forms a bell curve, with some types having benefited from more examples to learn from, while others suffered a decrease for reasons similar to those of "Is Functionally Equivalent To" and "Occurs in Genotype".

### 4.3.3 The impact of model transformations on Relation Extraction

Before the work on the transformation rules for SeeDev was completed, I used AlvisRE on the annotated corpus, using the annotation model directly. This excludes the Condition type, which takes another relation as an argument, as AlvisRE can only predict relations having two entity-arguments. These results (see Table 4.5) cannot be directly compared to the other results presented in this section, as they only include the data from pack1.

However, it is worth noting that results such as the ones obtained for “Is Found In or During” encouraged our consideration of producing finer grained models for RE. In spite of the fact that “Is Found In or During” was the type that was best represented in the data, it had a disproportionately low F-measure, due to the low prediction Precision. “Is Found In or During’ accounted for 29% of all relation annotations in pack 1 (see Table 4.5), and “Time and Localization” accounted for 20% of the relation occurrences in the seedev-binary dataset (see Table 2.3). With their distributions being similar, if the transformation had not improved the prediction quality, it would be expected that the scores for these two compared to the other

| **Type** | **# occ** | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|---|
| Belongs To | 33 | 0.319 | 0.455 | 0.375 |
| Binds To | 19 | 0.085 | 0.474 | 0.144 |
| Comparison | 55 | 0.355 | 0.691 | 0.469 |
| Interacts With | 1 | 0.000 | 0.000 | 0.000 |
| Is Found In or During | 139 | 0.199 | 0.475 | 0.280 |
| Encodes | 6 | 0.600 | 0.500 | 0.545 |
| Regulates Accumulation Of | 2 | 0.000 | 0.000 | 0.000 |
| Regulates Activity Of | 132 | 0.392 | 0.606 | 0.476 |
| Regulates Expression Of | 94 | 0.425 | 0.479 | 0.450 |
| **All relations** | **481** | **0.321** | **0.532** | **0.400** |

Table 4.5: Cross-Validation results using the annotation model on pack 1.

relation types would correspond before and after the transformation. In reality, while “Is Found In or During” had one of the worst F-measure scores, “Time and Localization” was the family of relation types that was best predicted by AlvisRE. Consequently, while it is not fair to directly compare the scores of Table 4.5 with the ones presented in Table 4.3, it is still safe to state that the decomposition of “Is Found In or During” into the category “Time and Localization” was beneficial to the quality of the prediction.

### 4.3.4 AlvisRE compared to the SeeDev participants

After using Cross-Validation to explore the impact of the margin parameter and relation types, I proceeded to use the online evaluation tool of the challenge[31] to compare my results to those obtained by the participants.

At this time, parameter optimization has not been fully performed for this task. As it was mentioned in the section 4.3.1, the margin parameter of the SVM plays a very important role in the performance of AlvisRE. Based on this and the promising results using cross-validation (table 4.3), I hope that AlvisRE can be optimized and achieve even better results on the challenge.

Based on the results shown in table 4.6, with an overall F-measure of 0.329, AlvisRE would be placed 6th in the challenge, after the ULISBOA team. For the various relation types, compared to the best F-measure obtained by the participants:

- Time and Localization: AlvisRE scores 2% lower
- Function: AlvisRE obtains an equivalent score
- Regulation: AlvisRE scores 6% lower
- Composition and Membership: AlvisRE scores 11.5% lower
- Interaction: AlvisRE scores 13.5% lower
- Is Linked To: AlvisRE fails to predict any relations and scores 15.4% lower, consequently.

In order to explore the use of semantic information on SeeDev, I used the vectors obtained by word2vec on the PubMed abstract corpus (see section 3.5.5). In this setup, the C parameter was

[31] http://2016.bionlp-st.org/tasks/seedev/seedev-evaluation

not optimized at all. In comparison, for the results obtained without word2vec (table 4.2), some optimization was performed, even though the process is not complete. Nevertheless, the results (seen in table 4.7) are promising, as using semantic information managed to improve AlvisRE's performance for some relation types :

- Time and Localization: the use of semantic information improved AlvisRE's score by 21.5%, placing it ahead of the participants by 21%.
- Function: scores improved by 8.3% and AvlisRE outperforms the best participant by 8.5%
- Regulation: scores are lower using semantic information
- Composition and Membership: improved by 2%, but remains outperformed by 11.5%
- Interaction: scores are lower using semantic information
- Is Linked To: AlvisRE fails to predict any relations in this case, too.

## 4.4 Discussion

Most active teams in the BioNLP community submitted their results for the SeeDev challenge, giving credibility to the corpus and a good measure for comparison for AlvisRE. The results obtained by the Seedev task participants were comparable to the results these methods obtain on other corpora. This constitutes empirical evidence that the SeeDev corpus is adapted to Information Extraction.

While I did not participate in the challenge, the results presented in this thesis show that AlvisRE is competitive and overall a good approach. On the seedev-binary task, it managed to achieve a good performance. Even though the results using cross-validation show great promise, those obtained using the online evaluation tool (see table 4.6) lack the appropriate parameter optimization and are, therefore, inconclusive regarding where AlvisRE is placed among the participants. Further experimentation will clarify both the comparison of AlvisRE to the participating teams, and the impact of semantic information on this task. It is still noteworthy that the highest ranking participant system, LitWay, was based on a rule-based approach, complimented by machine learning. By design, AlvisRE only includes machine learning approaches, based on the belief that even though hand-written rules may outperform ML-based systems in specific cases, they are not scalable nor adaptable. These approaches need to be re-developed for each new domain and their set of rules needs to be extended manually in case of model modifications.

Learning and predicting separately for each relation type (or class) yielded much better results, it introduced however a problematic behavior: for one pair of entities, multiple relation types may be predicted by AlvisRE, even though it is unlikely that one entity pair be in more than one relation at a time in the text. This is a limitation of AlvisRE, and every system that opts to treat classes separately[32], instead of applying a "1 vs all" classification method, which always picks one.

A phenomenon I observed while using AlvisRE on Seedev was that the number of candidates was very large. While this is understandable for a corpus of this size and a model of this complexity,

[32] The UniMelb system which participated in the seedev-binary challenge is another example.

| Relation Type | F-Measure | Precision | Recall |
|---|---|---|---|
| **All Relations** | **0.329** | **0.333** | **0.325** |
| **Time and Localization** | **0.120** | **0.112** | **0.129** |
| Exists At Stage | 0.182 | 0.130 | 0.300 |
| Exists In Genotype | 0.328 | 0.319 | 0.339 |
| Occurs During | 0.000 | 0.000 | 0.000 |
| Occurs In Genotype | 0.000 | 0.000 | 0.000 |
| Is Localized In | 0.336 | 0.280 | 0.419 |
| **Function** | **0.257** | **0.234** | **0.286** |
| Is Involved In Process | 0.000 | 0.000 | 0.000 |
| Transcribes Or Translates To | 0.333 | 0.357 | 0.313 |
| Is Functionally Equivalent To | 0.425 | 0.344 | 0.554 |
| **Regulation** | **0.358** | **0.385** | **0.335** |
| Regulates Accumulation | 0.125 | 0.125 | 0.125 |
| Regulates Development Phase | 0.284 | 0.296 | 0.273 |
| Regulates Expression | 0.335 | 0.361 | 0.312 |
| Regulates Molecule Activity | 0.000 | 0.000 | 0.000 |
| Regulates Process | 0.417 | 0.439 | 0.398 |
| Regulates Tissue Development | 0.000 | 0.000 | 0.000 |
| **Composition and Membership** | **0.375** | **0.357** | **0.394** |
| Composes Primary Structure | 0.333 | 0.500 | 0.250 |
| Composes Protein Complex | 0.000 | 0.000 | 0.000 |
| Has Sequence Identical To | 0.555 | 0.508 | 0.611 |
| Is Member Of Family | 0.310 | 0.287 | 0.337 |
| Is Protein Domain Of | 0.516 | 0.516 | 0.516 |
| **Interaction** | **0.178** | **0.197** | **0.163** |
| Interacts With | 0.125 | 0.121 | 0.130 |
| Binds To | 0.267 | 0.462 | 0.188 |
| **Specific to Binary Framework** | **0.000** | **0.000** | **0.000** |
| Is Linked To | 0.000 | 0.000 | 0.000 |

Table 4.6: The results of AlvisRE on the test set of seedev-binary, calculated by using the official evaluation tool.

| **Relation Type** | **F-Measure** | **Precision** | **Recall** |
|---|---:|---:|---:|
| **All Relations** | **0.283** | **0.213** | **0.319** |
| **Time and Localization** | **0.335** | **0.097** | **0.140** |
| Exists At Stage | 0.240 | 0.200 | 0.300 |
| Exists In Genotype | 0.388 | 0.363 | 0.417 |
| Occurs During | 0.000 | 0.000 | 0.000 |
| Occurs In Genotype | 0.000 | 0.000 | 0.000 |
| Is Localized In | 0.371 | 0.315 | 0.452 |
| **Function** | **0.340** | **0.289** | **0.422** |
| Is Involved In Process | 0.000 | 0.000 | 0.000 |
| Transcribes Or Translates To | 0.313 | 0.313 | 0.313 |
| Is Functionally Equivalent To | 0.432 | 0.352 | 0.561 |
| **Regulation** | **0.255** | **0.211** | **0.396** |
| Regulates Accumulation | 0.000 | 0.000 | 0.000 |
| Regulates Development Phase | 0.222 | 0.224 | 0.221 |
| Regulates Expression | 0.356 | 0.358 | 0.355 |
| Regulates Molecule Activity | 0.000 | 0.000 | 0.000 |
| Regulates Process | 0.243 | 0.160 | 0.505 |
| Regulates Tissue Development | 0.000 | 0.000 | 0.000 |
| **Composition and Membership** | **0.395** | **0.375** | **0.424** |
| Composes Primary Structure | 0.467 | 0.500 | 0.438 |
| Composes Protein Complex | 0.000 | 0.000 | 0.000 |
| Has Sequence Identical To | 0.508 | 0.432 | 0.615 |
| Is Member Of Family | 0.287 | 0.274 | 0.302 |
| Is Protein Domain Of | 0.557 | 0.567 | 0.548 |
| **Interaction** | **0.159** | **0.209** | **0.137** |
| Interacts With | 0.106 | 0.102 | 0.111 |
| Binds To | 0.255 | 0.400 | 0.188 |
| **Specific to Binary Framework** | **0.000** | **0.000** | **0.000** |
| Is Linked To | 0.000 | 0.000 | 0.000 |

Table 4.7: The results of AlvisRE with word2vec semantic similarities on the test set of seedev-binary, calculated by using the official evaluation tool.

more than 99% of these candidates were false. This led to a waste in computation resources. This shortcoming might not be unique to AlvisRE, as the model creates lots of possible relations even by simple co-occurrence, but it is an issue that will need to be addressed when dealing with large corpora and complex models that create multiple possibilities for each pair of arguments.

While developing AlvisRE I set in place a great number of variables and parameters that could be tweaked to increase performance. As it is often the case with software parameter choice, this makes optimization costly. Fortunately, as AlvisRE matures as a software, good values are pre-selected by default, leaving the choice of further tweaking optional, for the advanced user. Furthermore, learning strategies adapted to complex IE tasks can be devised, such as parameter optimization on smaller subsets in order to cut down on computation time.

# Conclusion & Future Work

**4.1 Introduction** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **119**
**4.2 SeeDev in BioNLP-ST '16** . . . . . . . . . . . . . . . . . . . . . . . . . . **120**
**4.3 AlvisRE on *Arabidopsis thaliana*** . . . . . . . . . . . . . . . . . . . . . . **122**
4.3.1 Optimizing the SVM margin parameter . . . . . . . . . . . . . . . . . . . 122
4.3.2 Exploring different relation types . . . . . . . . . . . . . . . . . . . . . . . 122
4.3.3 The impact of model transformations on Relation Extraction . . . . . . 124
4.3.4 AlvisRE compared to the SeeDev participants . . . . . . . . . . . . . . . 125
**4.4 Discussion** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **126**

## 1 Conclusion

This thesis presented a complete IE task on *A. thaliana*, a model plant, with the goal of showing how Information Extraction can deliver structured, usable information for real-world problems in biology. Two axes were defined throughout this study: one regarding the knowledge model and data, and a second one concerning the algorithmic approach for Relation Extraction. These two axes are interconnected and were explored in an iterative way, with the findings on one influencing choices on the other.

The complexity of the model and data that can capture the knowledge of a biological domain such as this one are greater than the ones usually seen in past IE tasks. By working closely with a multidisciplinary team of experts, I produced models which adapt to the contradicting needs of accurate knowledge representation and enough generalisation for machine learning. The conceptual model presented in this thesis adequately represents the knowledge structure for regulatory networks in seed development for *A. thaliana* and it can be reused for other applications, such as the building of ontologies.

My goal for the relation extraction aspect of this study has been to produce an algorithmic approach that can optimize the use of different types of knowledge in order to generalize best and increase performance in complex tasks and also be easily adaptable to new tasks. By using different types of information in a modular way, the Relation Extraction approach introduced in this thesis, AlvisRE, can perform well in different, complex tasks. This approach takes advantage of syntax and semantics in order to generalize easily, and consequently it has good results on smaller sets of training data.

AlvisRE managed to perform better than the current state of the art in all the scenarios tested. These scenarios covered different sizes of datasets, types of biological information and model complexity. However, they were all focused on biology. A reasonable next step would be using AlvisRE on medical data, a closely related field, but one which has different challenges.

The annotated data served their primary goal of training a machine learning model for IE on the subject. Nevertheless, their use is not limited to the scope of this work, as the resulting learning model can be used to extract more information, and the extracted information can be used for future applications for Systems Biology such as visualization of the regulatory networks, population of databases, semantic search engines, etc. The importance of such applications is great, as Systems Biology serves to describe, predict and fully exploit biological knowledge. As a matter of fact, the SeeDev use case has been chosen as part of the OpenMinted project[33], where the data extracted thanks to this study will be integrated in the FLAGdb[34] [Dérozier et al., 2011] database of the Institute of Plant Sciences Paris-Saclay[35].

Additionally the model and the data were used in the recent international challenge in biomedical IE, BioNLP ST '16, and they defined a new subtask in this recurring shared task: Seedev[36]. They will without a doubt continue to be a part of future editions of BioNLP.

Furthermore, the corpus and model will be directly used in a collaboration with the BASE team, led by Christine Dillmann from the INRA lab “GQE Le Moulon”[37]. The next step in this collaboration will be extracting relations from texts on *Z. mays* (maize), using the learning models trained on the *A. Thaliana* corpus, and Named Entities extracted thanks to the relevant nomenclature.

The goal of this thesis has been since the beginning to successfully perform a full IE task. While it has succeeded in this goal, there were a number of limitations. Within the limits of this work, the extracted data were never used in an application (like visualization), in order to have the experts evaluate this approach as a part of their process. The usage possibilities remain to be explored and evaluated in the future.

Even though it is my ambition that the models trained on *A. thaliana* using this corpus can be used directly on other plants, this has not been tested. At the current stage, the expert remains necessary to pre-annotate data. Moreover, the models produced for data representation have not been evaluated, with or without modification, on other organisms.

Text is an alternative source of information, but other sources such as experimental data, and manually curated databases already exist. These sources could be combined with an IE approach. While this complementarity was taken into account while working on this thesis, data from other sources were not used. Their integration into this approach could take place as a source of semantic information, or as a source for distant supervision for the learning approach.

Finally, this work has shown that computation time can be a true limitation in the scale of real-world datasets. Optimizations are necessary in both algorithm development but also in

[33] http://openminted.eu/

[34] http://tools.ips2.u-psud.fr/projects/FLAGdb++/HTML/index.shtml

[35] http://www.ips2.u-psud.fr/?lang=fr

[36] http://2016.bionlp-st.org/tasks/seedev

[37] http://moulon.inra.fr/

learning strategy. AlvisRE was programmed with multi-threading capabilities in order to alleviate the cost of computation, but the processing of large sets of data can still necessitate several weeks of CPU time. Such limitations make the use of more complex algorithms on large volumes of data restricted in the industry, and with IE exiting the context of benchmark challenges into the world of applied science, optimizing computation time will need to be central in future research.

Information Extraction plays an increasingly important role in biology. Therefore, a successful study of the full cycle of an IE task serves as a solid example for both scientific communities. Additionally, working with plant seeds which represent a great stake for the agricultural industry, the topic is guaranteed to continue to attract interest, with this work laying the foundation for a number of future projects.

## 2 Future Work

### 2.1 Data

Thanks to the continuity offered by current and future projects like the BioNLP Shared Task and OpenMinTeD, the **annotation** and publication of the rest of the packs is already underway and it will be completed in the next years.

The FLAGdb[38]-based application in collaboration with IPS2[39] in the context of project OpenMinTeD[40] (use case SeeDev) will be the first step into putting the results of this work into application. Other **applications** envisioned are the develpment of visualizations of the resulting regulatory networks, the combination with other sources, such as experimental data.

In the same spirit of extending this foundation work, the models trained for extraction in the context of this work will be used on **new, unannotated data** in order to enrich the extracted information. Once this information reaches a level where it can offer valuable insight for the *A. thaliana* community, the extracted data and their transformations can be the subject of new scientific publications.

As far as the model is concerned, making it available in the form of an ontology in the IBC AgroPortal[41] and the NCBO BioPortal[42] would be the next step in the direction of dissemination of information.

### 2.2 Information Extraction

The next step for AlvisRE will be **event extraction**. Models such as the ones described here focus on n-ary events. Most contemporary RE systems, like AlvisRE, extract binary relations which are later used to construct more complex events. It was in this spirit, for example, that the BioNLP '16 SeeDev challenge[43] offered two versions of the RE task: seedev-full and seedev-binary, corresponding to n-ary event and binary relation extraction, respectively.

[38] [Dérozier et al., 2011]

[39] http://www.ips2.u-psud.fr/

[40] http://openminted.eu/

[41] http://agroportal.lirmm.fr/

[42] http://bioportal.bioontology.org/

[43] http://2016.bionlp-st.org/tasks/seedev

Additionally, even though **modalities** such as negation and speculation were annotated by the experts, they were never used in either SeeDev sub-tasks, nor were they treated by AlvisRE. As these relation modifiers play an important role semantically, it would be necessary to extend AvisRE to cover them.

In the context of this work, I studied the impact of using semantic information and found **distributional semantics** approaches promising. The developments in distributional semantics have been great in the past few years, and since the appearance of word2vec a number of other approaches have appeared, like GloVe from Stanford NLP [Pennington et al., 2014]. GloVe has been shown to have comparable performances to word2vec in benchmarks. AlvisRE has been tested with word2vec, but GloVe or other tools could be easily integrated in the architecture. Approaches which can use different contexts than the windows of word2vec are also of particular interest, as they allow the use of contexts such as the dependency graph. An example of such an approach is the word2vec reimplementation by Goldberg and Levy, word2vecf [Levy and Goldberg, 2014]. Using dependencies would necessitate the pre-processing of large corpora and thus reduce the efficiency, but with a promise of improvement of the word representations, such scenarios would possibly require smaller corpora for equivalent quality of results.

The **syntactic information** used with AlvisRE at this moment was parsings using the CCG parser [Clark and Curran, 2007] and AlvisGrammar [Ratkovic, 2014, Ratkovic et al., 2012]. Other dependency parsers like the recent neural network Stanford Parser [Chen and Manning, 2014] could be tested, as they correspond to the current state-of-the-art in dependency parsers. Additionally, AlvisRE could be made to work with constituent parsers.

As far as the **machine learning** core of AlvisRE is concerned, there are a number of possibilities for future work. Kernel stacking and other multi-kernel approaches [Miwa et al., 2009, He et al., 2013] have been shown to outperform single-kernel methods in recent works. Another possibility would be removing the EKM transformation and use non-SVD kernels directly like in the works of [Moschitti and Zanzotto, 2007, Boughorbel et al., 2004, Neuhaus and Bunke, 2006, Haasdonk, 2005, Hsuan-tien Lin, 2003]. These kernels might not satisfy the mathematical properties of kernels, but have been empirically shown to work well, with the risk of using them being finding local minima instead of an optimal solution. On the other extreme of the replacement of EKM, another idea worth exploring is using more features. While I focused on kernel-based approaches in my study, feature-based systems have been long used in RE, and this literature could prove a valuable source for the improvement of AlvisRE.

A subject addressed earlier in this chapter is that of *computation time*, a problem that becomes relevant when moving from small datasets to real-world volumes. An extensive study of solutions to this problem would include carefully crafted learning strategies and a deep analysis of the algorithms and code used. The efficiency of EKM and Global Alignment has been discussed in other works such as [Liao and Noble, 2002, Liao and Noble, 2003, Schölkopf et al., 2004], and notably Scholkopf et al describe ways to alleviate the computational cost of the EKM transformation.

Furthering the reuse of AlvisRE, its full integration in the Alvis pipeline is planned for the future. This would also enable AlvisRE to be a part of project OpenMinted which aims to facilitate the reuse of various IE systems, by assuring their interoperability.

Finally, I believe that it would be worthwhile to compare and test things learned within the context of RE in *other use cases*. For example, sentence similarity is a cornerstone part of many other domains, such as textual entailment, text classification, natural language understanding, *etc.* A Global Alignment approach such as the one used here could prove to be appropriate to these related fields, in its entirety or partly.

# Appendices

# Appendix A

# Manual Classes for LLL

This set includes 22 classes of a size varying from two to 18 words and they were created for the LLL/BI corpus.

The manual classes used here are the following:

1. transcription, expression, activity, assembly, sequestration, synthesis, phosphorylation, dephosphorylation, processing
2. regulate, control, activate, inhibit, block, limit, induce, affect, drive, repress, transcribe, increase, decrease, stimulate, contribute, direct, cause, influence
3. require, need
4. activator, regulator, inhibitor, repressor, represser
5. activation, inhibition, regulation, control, production, increase, decrease, action, repression, induction
6. essential, required, necessary, responsible, sufficient, needed
7. dependent, dependant, sensitive
8. depend, rely
9. protein, factor, product, kinase, phosphatase, enzyme, polymerase, holoenzyme, transferase
10. capacity, ability
11. member, part,
12. manner, way, fashion
13. recognize, bind, adhere, use
14. region, promoter, element, sequence
15. precede, bind, span, recognize
16. gene, operon
17. level, presence, concentration

18. function, sporulation, growth, stage, formation, division, outgrow
19. starvation, stress
20. prespore, endospore, forespore, compartment
21. upstream, downstream
22. demonstrate, show, indicate, suggest, conclude, believe

# Appendix B

# Trigger words for Bacteria Biotopes

- parasit
- attack
- coloniz
- flora
- infect
- inhabit
- invade
- important
- host
- environment
- niche
- habitat
- effect
- contamin
- ecolog
- toward
- presen
- subject
- induce
- implicate
- ingest
- grow
- detect
- found
- live
- spread
- survive
- unable
- commensal
- isolate
- symbio
- relationship
- present
- discover
- observ
- econom
- disease
- virulence
- chronic
- symptom
- syndrome
- severe
- fever
- caus
- treat
- prevalence
- outbreak
- epidem
- ill
- pathogen
- phytopathogen
- infest
- ingest
- fed
- eat

# Appendix C

# A complete AlvisRE input file

This appendix contains an entire training input file for Alvis RE, from the Bacteria Biotope corpus.

T1-BTID-10098 Sentence 0 21 Brucella abortus S19
T2-BTID-10098 Sentence 21 33 Description
T3-BTID-10098 Sentence 33 42 Brucella.
T4-BTID-10098 Sentence 43 183 There are 7 Brucella spp., of which four are pathogenic in humans (Brucella melitensis, Brucella abortus, Brucella canis and Brucella suis).
T5-BTID-10098 Sentence 184 339 They are highly infectious, and can be spread through contact with infected animal products or through the air, making them a potential bioterrorism agent.
T6-BTID-10098 Sentence 340 558 Once the organism has entered the body, it can become intracellular, and enter the blood and lymphatic regions, multiplying inside phagocytes before eventually causing bacteremia (spread of bacteria through the blood).
T7-BTID-10098 Sentence 559 692 Virulence may depend on a type IV secretion system which may promote intracellular growth by secreting important effector molecules.
T8-BTID-10098 Sentence 692 704 Description
T9-BTID-10098 Sentence 704 721 Brucella abortus.
T10-BTID-10098 Sentence 722 783 This organism was first noticed on the island of Malta by Dr.
T11-BTID-10098 Sentence 784 838 David Bruce during an epidemic among British soldiers.
T12-BTID-10098 Sentence 839 1004 It is the primary cause of bovine brucellosis, which results in enormous (billions of dollars) economic losses due primarily to reproductive failure and food losses.
T13-BTID-10098 Sentence 1005 1127 In man, it causes undulant fever, a long debilitating disease that is treated by protracted administration of antibiotics.
T14-BTID-10098 Sentence 1128 1215 Brucella abortus is listed as a civilian, military, and agricultural bioterrism agent.
T15-BTID-10098 Sentence 1215 1227 Description
T16-BTID-10098 Sentence 1227 1248 Brucella abortus S19.
T17-BTID-10098 Sentence 1249 1308 This is a spontaneously attenuated strain discovered by Dr.
T18-BTID-10098 Sentence 1309 1327 John Buck in 1923.
T19-BTID-10098 Sentence 1328 1443 However, the underlying molecular or physiological mechanisms causing the loss of virulence is not well understood.
T20-BTID-10098 Sentence 1444 1610 Since early 1930s, this strain has been used worldwide as

```
an effective vaccine to prevent brucellosis in cattle until it was replaced by strain RB51 during
the 1990s.
T21-BTID-10098 Sentence 1611 1789 The main objective of this project is to identify genes
associated with the virulence or lack there of, through comparison of S19 genome with that of
the virulent counterparts.
T1 Word 0 20 Brucella abortus S19|Brucella_abortus_S19|NN
T2 Word 20 21 |_|.
T3 Word 21 32 Description|Description|NN
T4 Word 32 33 |_|.
T5 Word 33 41 Brucella|Brucella|NN
T6 Word 41 42 .|.|.
T7 Word 43 48 There|There|EX
T8 Word 49 52 are|be|VBP
T9 Word 53 54 7|7|CD
T10 Word 55 68 Brucella spp.|Brucella_spp.|NNP
T11 Word 68 69 ,|,|,
T12 Word 70 72 of|of|IN
T13 Word 73 78 which|which|WDT
T14 Word 79 83 four|four|CD
T15 Word 84 87 are|be|VBP
T16 Word 88 98 pathogenic|pathogenic|JJ
T17 Word 99 101 in|in|IN
T18 Word 102 108 humans|human|NNS
T19 Word 109 110 (|(|LRB
T20 Word 110 129 Brucella melitensis|Brucella_melitensis|NN
T21 Word 129 130 ,|,|,
T22 Word 131 147 Brucella abortus|Brucella_abortus|NN
T23 Word 147 148 ,|,|,
T24 Word 149 163 Brucella canis|Brucella_canis|NN
T25 Word 164 167 and|and|CC
T26 Word 168 181 Brucella suis|Brucella_suis|NN
T27 Word 181 182 )|)|RRB
T28 Word 182 183 .|.|.
T29 Word 184 188 They|They|PRP
T30 Word 189 192 are|be|VBP
T31 Word 193 199 highly|highly|RB
T32 Word 200 210 infectious|infectious|JJ
T33 Word 210 211 ,|,|,
T34 Word 212 215 and|and|CC
T35 Word 216 219 can|can|MD
T36 Word 220 222 be|be|VB
T37 Word 223 229 spread|spread|VBN
T38 Word 230 237 through|through|IN
T39 Word 238 245 contact|contact|NN
T40 Word 246 250 with|with|IN
T41 Word 251 259 infected|infected|JJ
T42 Word 260 266 animal|animal|NN
T43 Word 267 275 products|product|NNS
```

```
T44 Word 276 278 or|or|CC
T45 Word 279 286 through|through|IN
T46 Word 287 290 the|the|DT
T47 Word 291 294 air|air|NN
T48 Word 294 295 ,|,|,
T49 Word 296 302 making|make|VBG
T50 Word 303 307 them|them|PRP
T51 Word 308 309 a|a|DT
T52 Word 310 319 potential|potential|JJ
T53 Word 320 332 bioterrorism|bioterrorism|NN
T54 Word 333 338 agent|agent|NN
T55 Word 338 339 .|.|.
T56 Word 340 344 Once|Once|IN
T57 Word 345 348 the|the|DT
T58 Word 349 357 organism|organism|NN
T59 Word 358 361 has|have|VBZ
T60 Word 362 369 entered|enter|VBD
T61 Word 370 373 the|the|DT
T62 Word 374 378 body|body|NN
T63 Word 378 379 ,|,|,
T64 Word 380 382 it|it|PRP
T65 Word 383 386 can|can|MD
T66 Word 387 393 become|become|VB
T67 Word 394 407 intracellular|intracellular|JJ
T68 Word 407 408 ,|,|,
T69 Word 409 412 and|and|CC
T70 Word 413 418 enter|enter|VBP
T71 Word 419 422 the|the|DT
T72 Word 423 428 blood|blood|NN
T73 Word 429 432 and|and|CC
T74 Word 433 442 lymphatic|lymphatic|JJ
T75 Word 443 450 regions|region|NNS
T76 Word 450 451 ,|,|,
T77 Word 452 463 multiplying|multiply|VBG
T78 Word 464 470 inside|inside|JJ
T79 Word 471 481 phagocytes|phagocyte|NNS
T80 Word 482 488 before|before|IN
T81 Word 489 499 eventually|eventually|RB
T82 Word 500 507 causing|cause|VBG
T83 Word 508 518 bacteremia|bacteremia|NN
T84 Word 519 520 (|(|LRB
T85 Word 520 526 spread|spread|NN
T86 Word 527 529 of|of|IN
T87 Word 530 538 bacteria|bacteria|NNS
T88 Word 539 546 through|through|IN
T89 Word 547 550 the|the|DT
T90 Word 551 556 blood|blood|NN
T91 Word 556 557 )|)|RRB
```

```
T92 Word 557 558 .|.|.
T93 Word 559 568 Virulence|Virulence|NN
T94 Word 569 572 may|may|MD
T95 Word 573 579 depend|depend|VB
T96 Word 580 582 on|on|IN
T97 Word 583 584 a|a|DT
T98 Word 585 589 type|type|NN
T99 Word 590 592 IV|IV|CD
T100 Word 593 602 secretion|secretion|NN
T101 Word 603 609 system|system|NN
T102 Word 610 615 which|which|WDT
T103 Word 616 619 may|may|MD
T104 Word 620 627 promote|promote|VB
T105 Word 628 641 intracellular|intracellular|JJ
T106 Word 642 648 growth|growth|NN
T107 Word 649 651 by|by|IN
T108 Word 652 661 secreting|secrete|VBG
T109 Word 662 671 important|important|JJ
T110 Word 672 680 effector|effector|NN
T111 Word 681 690 molecules|molecule|NNS
T112 Word 690 691 .|.|.
T113 Word 691 692 |_|.
T114 Word 692 703 Description|Description|NN
T115 Word 703 704 |_|.
T116 Word 704 720 Brucella abortus|Brucella_abortus|NN
T117 Word 720 721 .|.|.
T118 Word 722 726 This|This|DT
T119 Word 727 735 organism|organism|NN
T120 Word 736 739 was|be|VBD
T121 Word 740 745 first|first|JJ
T122 Word 746 753 noticed|notice|VBN
T123 Word 754 756 on|on|IN
T124 Word 757 760 the|the|DT
T125 Word 761 767 island|island|NN
T126 Word 768 770 of|of|IN
T127 Word 771 776 Malta|Malta|NN
T128 Word 777 779 by|by|IN
T129 Word 780 782 Dr|Dr|NN
T130 Word 782 783 .|.|.
T131 Word 784 789 David|David|NN
T132 Word 790 795 Bruce|Bruce|NN
T133 Word 796 802 during|during|IN
T134 Word 803 805 an|an|DT
T135 Word 806 814 epidemic|epidemic|JJ
T136 Word 815 820 among|among|IN
T137 Word 821 828 British|British|NN
T138 Word 829 837 soldiers|soldier|NNS
T139 Word 837 838 .|.|.
```

```
T140 Word 839 841 It|It|PRP
T141 Word 842 844 is|be|VBZ
T142 Word 845 848 the|the|DT
T143 Word 849 856 primary|primary|JJ
T144 Word 857 862 cause|cause|NN
T145 Word 863 865 of|of|IN
T146 Word 866 872 bovine|bovine|JJ
T147 Word 873 884 brucellosis|brucellosis|NN
T148 Word 884 885 ,|,|,
T149 Word 886 891 which|which|WDT
T150 Word 892 899 results|result|VBZ
T151 Word 900 902 in|in|IN
T152 Word 903 911 enormous|enormous|JJ
T153 Word 912 913 (|(|LRB
T154 Word 913 921 billions|billion|NNS
T155 Word 922 924 of|of|IN
T156 Word 925 932 dollars|dollar|NNS
T157 Word 932 933 )|)|RRB
T158 Word 934 942 economic|economic|JJ
T159 Word 943 949 losses|loss|NNS
T160 Word 950 953 due|due|JJ
T161 Word 954 963 primarily|primarily|RB
T162 Word 964 966 to|to|TO
T163 Word 967 979 reproductive|reproductive|JJ
T164 Word 980 987 failure|failure|NN
T165 Word 988 991 and|and|CC
T166 Word 992 996 food|food|NN
T167 Word 997 1003 losses|loss|NNS
T168 Word 1003 1004 .|.|.
T169 Word 1005 1007 In|In|IN
T170 Word 1008 1011 man|man|NN
T171 Word 1011 1012 ,|,|,
T172 Word 1013 1015 it|it|PRP
T173 Word 1016 1022 causes|cause|VBZ
T174 Word 1023 1031 undulant|undulant|JJ
T175 Word 1032 1037 fever|fever|NN
T176 Word 1037 1038 ,|,|,
T177 Word 1039 1040 a|a|DT
T178 Word 1041 1045 long|long|JJ
T179 Word 1046 1058 debilitating|debilitating|JJ
T180 Word 1059 1066 disease|disease|NN
T181 Word 1067 1071 that|that|WDT
T182 Word 1072 1074 is|be|VBZ
T183 Word 1075 1082 treated|treat|VBN
T184 Word 1083 1085 by|by|IN
T185 Word 1086 1096 protracted|protracted|JJ
T186 Word 1097 1111 administration|administration|NN
T187 Word 1112 1114 of|of|IN
```

```
T188 Word 1115 1126 antibiotics|antibiotic|NNS
T189 Word 1126 1127 .|.|.
T190 Word 1128 1144 Brucella
abortus|Brucella_abortus|NN
T191 Word 1145 1147 is|be|VBZ
T192 Word 1148 1154 listed|list|VBN
T193 Word 1155 1157 as|as|IN
T194 Word 1158 1159 a|a|DT
T195 Word 1160 1168 civilian|civilian|JJ
T196 Word 1168 1169 ,|,|,
T197 Word 1170 1178 military|military|JJ
T198 Word 1178 1179 ,|,|,
T199 Word 1180 1183 and|and|CC
T200 Word 1184 1196 agricultural|agricultural|JJ
T201 Word 1197 1207 bioterrism|bioterrism|NN
T202 Word 1208 1213 agent|agent|NN
T203 Word 1213 1214 .|.|.
T204 Word 1214 1215 |_|.
T205 Word 1215 1226 Description|Description|NN
T206 Word 1226 1227 |_|.
T207 Word 1227 1247 Brucella abortus S19|Brucella_abortus_S19|NN
T208 Word 1247 1248 .|.|.
T209 Word 1249 1253 This|This|DT
T210 Word 1254 1256 is|be|VBZ
T211 Word 1257 1258 a|a|DT
T212 Word 1259 1272 spontaneously|spontaneously|RB
T213 Word 1273 1283 attenuated|attenuate|VBN
T214 Word 1284 1290 strain|strain|NN
T215 Word 1291 1301 discovered|discover|VBN
T216 Word 1302 1304 by|by|IN
T217 Word 1305 1307 Dr|Dr|NN
T218 Word 1307 1308 .|.|.
T219 Word 1309 1313 John|John|NN
T220 Word 1314 1318 Buck|Buck|NN
T221 Word 1319 1321 in|in|IN
T222 Word 1322 1326 1923|1923|CD
T223 Word 1326 1327 .|.|.
T224 Word 1328 1335 However|However|RB
T225 Word 1335 1336 ,|,|,
T226 Word 1337 1340 the|the|DT
T227 Word 1341 1351 underlying|underlie|JJ
T228 Word 1352 1361 molecular|molecular|JJ
T229 Word 1362 1364 or|or|CC
T230 Word 1365 1378 physiological|physiological|JJ
T231 Word 1379 1389 mechanisms|mechanism|NNS
T232 Word 1390 1397 causing|cause|VBG
T233 Word 1398 1401 the|the|DT
T234 Word 1402 1406 loss|loss|NN
```

```
T235 Word 1407 1409 of|of|IN
T236 Word 1410 1419 virulence|virulence|NN
T237 Word 1420 1422 is|be|VBZ
T238 Word 1423 1426 not|not|RB
T239 Word 1427 1431 well|well|RB
T240 Word 1432 1442 understood|understand|VBN
T241 Word 1442 1443 .|.|.
T242 Word 1444 1449 Since|Since|IN
T243 Word 1450 1455 early|early|JJ
T244 Word 1456 1461 1930s|1930|NNS
T245 Word 1461 1462 ,|,|,
T246 Word 1463 1467 this|this|DT
T247 Word 1468 1474 strain|strain|NN
T248 Word 1475 1478 has|have|VBZ
T249 Word 1479 1483 been|be|VBN
T250 Word 1484 1488 used|use|VBN
T251 Word 1489 1498 worldwide|worldwide|NN
T252 Word 1499 1501 as|as|IN
T253 Word 1502 1504 an|an|DT
T254 Word 1505 1514 effective|effective|JJ
T255 Word 1515 1522 vaccine|vaccine|NN
T256 Word 1523 1525 to|to|TO
T257 Word 1526 1533 prevent|prevent|VB
T258 Word 1534 1545 brucellosis|brucellosis|NN
T259 Word 1546 1548 in|in|IN
T260 Word 1549 1555 cattle|cattle|NN
T261 Word 1556 1561 until|until|IN
T262 Word 1562 1564 it|it|PRP
T263 Word 1565 1568 was|be|VBD
T264 Word 1569 1577 replaced|replace|VBN
T265 Word 1578 1580 by|by|IN
T266 Word 1581 1587 strain|strain|NN
T267 Word 1588 1592 RB51|RB51|NN
T268 Word 1593 1599 during|during|IN
T269 Word 1600 1603 the|the|DT
T270 Word 1604 1609 1990s|1990s|NN
T271 Word 1609 1610 .|.|.
T272 Word 1611 1614 The|The|DT
T273 Word 1615 1619 main|main|JJ
T274 Word 1620 1629 objective|objective|NN
T275 Word 1630 1632 of|of|IN
T276 Word 1633 1637 this|this|DT
T277 Word 1638 1645 project|project|NN
T278 Word 1646 1648 is|be|VBZ
T279 Word 1649 1651 to|to|TO
T280 Word 1652 1660 identify|identify|VB
T281 Word 1661 1666 genes|gene|NNS
T282 Word 1667 1677 associated|associate|VBN
```

```
T283 Word 1678 1682 with|with|IN
T284 Word 1683 1686 the|the|DT
T285 Word 1687 1696 virulence|virulence|NN
T286 Word 1697 1699 or|or|CC
T287 Word 1700 1704 lack|lack|VBP
T288 Word 1705 1710 there|there|EX
T289 Word 1711 1713 of|of|IN
T290 Word 1713 1714 ,|,|,
T291 Word 1715 1722 through|through|IN
T292 Word 1723 1733 comparison|comparison|NN
T293 Word 1734 1736 of|of|IN
T294 Word 1737 1740 S19|S19|NN
T295 Word 1741 1747 genome|genome|NN
T296 Word 1748 1752 with|with|IN
T297 Word 1753 1757 that|that|DT
T298 Word 1758 1760 of|of|IN
T299 Word 1761 1764 the|the|DT
T300 Word 1765 1773 virulent|virulent|JJ
T301 Word 1774 1786 counterparts|counterpart|NNS
T302 Word 1786 1787 .|.|.
T303 Word 1787 1788 |_|.
T304 Word 1788 1789 |_|.
T305 Bacteria 0 20 Brucella abortus S19|Brucella_abortus_S19
T306 Bacteria 33 41 Brucella|Brucella
T307 Bacteria 55 68 Brucella spp.|Brucella_spp.
T308 Habitat 102 108 humans|humans
T309 Bacteria 110 129 Brucella melitensis|Brucella_melitensis
T310 Bacteria 131 147 Brucella abortus|Brucella_abortus
T311 Bacteria 149 163 Brucella canis|Brucella_canis
T312 Bacteria 168 181 Brucella suis|Brucella_suis
T313 Habitat 260 275 animal products|animal products
T314 Habitat 291 294 air|air
T315 Habitat 374 378 body|body
T316 Habitat 394 407 intracellular|intracellular
T317 Habitat 423 428 blood|blood
T318 Habitat 433 442 lymphatic|lymphatic
T319 Habitat 471 481 phagocytes|phagocytes
T320 Bacteria 530 538 bacteria|bacteria
T321 Habitat 551 556 blood|blood
T322 Habitat 628 641 intracellular|intracellular
T323 Bacteria 704 720 Brucella abortus|Brucella_abortus
T324 Geographical 761 776 island of Malta|island of Malta
T325 Habitat 821 837 British soldiers|British soldiers
T326 Geographical 821 828 British|British
T327 Habitat 866 872 bovine|bovine
T328 Habitat 992 996 food|food
T329 Habitat 1008 1011 man|man
T330 Bacteria 1128 1144 Brucella abortus|Brucella_abortus
```

```
T331 Bacteria 1227 1247 Brucella abortus S19|Brucella_abortus_S19
T332 Habitat 1549 1555 cattle|cattle
T333 Bacteria 1588 1592 RB51|RB51
T334 Bacteria 1737 1740 S19|S19
R1 Dependency dependent:T9 sentence:T4-BTID-10098 head:T10 label:NUM:N-N
R2 Dependency dependent:T16 sentence:T4-BTID-10098 head:T15 label:XCOMP:V-ADJ
R3 Dependency dependent:T24 sentence:T4-BTID-10098 head:T26 label:COORD_and:N-N
R4 Dependency dependent:T22 sentence:T4-BTID-10098 head:T26 label:COORD_and:N-N
R5 Dependency dependent:T20 sentence:T4-BTID-10098 head:T26 label:COORD_and:N-N
R6 Dependency dependent:T18 sentence:T4-BTID-10098 head:T26 label:APPOS:N-N
R7 Dependency dependent:T18 sentence:T4-BTID-10098 head:T24 label:APPOS:N-N
R8 Dependency dependent:T18 sentence:T4-BTID-10098 head:T22 label:APPOS:N-N
R9 Dependency dependent:T18 sentence:T4-BTID-10098 head:T20 label:APPOS:N-N
R10 Dependency dependent:T18 sentence:T4-BTID-10098 head:T15 label:COMP_in:V-N
R11 Dependency dependent:T14 sentence:T4-BTID-10098 head:T15 label:SUBJ:V-N
R12 Dependency dependent:T10 sentence:T4-BTID-10098 head:T15 label:SUBJ:V-N
R13 Dependency dependent:T10 sentence:T4-BTID-10098 head:T8 label:XCOMP:V-N
R14 Dependency dependent:T31 sentence:T5-BTID-10098 head:T32 label:MOD:ADJ-ADV
R15 Dependency dependent:T32 sentence:T5-BTID-10098 head:T30 label:XCOMP:V-ADJ
R16 Dependency dependent:T42 sentence:T5-BTID-10098 head:T43 label:MOD_ATT:N-N
R17 Dependency dependent:T41 sentence:T5-BTID-10098 head:T43 label:MOD_ATT:N-ADJ
R18 Dependency dependent:T43 sentence:T5-BTID-10098 head:T39 label:COMP_with:N-N
R19 Dependency dependent:T39 sentence:T5-BTID-10098 head:T37 label:COMP_through:V-N
R20 Dependency dependent:T47 sentence:T5-BTID-10098 head:T37 label:COMP_through:V-N
R21 Dependency dependent:T38 sentence:T5-BTID-10098 head:T45 label:COORD_or:IN-IN
R22 Dependency dependent:T36 sentence:T5-BTID-10098 head:T37 label:AUX:V-V
R23 Dependency dependent:T50 sentence:T5-BTID-10098 head:T49 label:OBJ:V-N
R24 Dependency dependent:T53 sentence:T5-BTID-10098 head:T54 label:MOD_ATT:N-N
R25 Dependency dependent:T52 sentence:T5-BTID-10098 head:T54 label:MOD_ATT:N-ADJ
R26 Dependency dependent:T54 sentence:T5-BTID-10098 head:T49 label:OBJ:V-N
R27 Dependency dependent:T49 sentence:T5-BTID-10098 head:T37 label:XMOD:V_PASS-V
R28 Dependency dependent:T35 sentence:T5-BTID-10098 head:T37 label:AUX:V-V
R29 Dependency dependent:T30 sentence:T5-BTID-10098 head:T37 label:COORD_and:V_PASS-V
R30 Dependency dependent:T29 sentence:T5-BTID-10098 head:T30 label:SUBJ:V-N
R31 Dependency dependent:T29 sentence:T5-BTID-10098 head:T37 label:SUBJ:V_PASS-N
R32 Dependency dependent:T62 sentence:T6-BTID-10098 head:T60 label:OBJ:V-N
R33 Dependency dependent:T59 sentence:T6-BTID-10098 head:T60 label:AUX:V-V
R34 Dependency dependent:T58 sentence:T6-BTID-10098 head:T60 label:SUBJ:V-N
R35 Dependency dependent:T60 sentence:T6-BTID-10098 head:T70 label:COMP_once:V-V
R36 Dependency dependent:T65 sentence:T6-BTID-10098 head:T66 label:AUX:V-V
R37 Dependency dependent:T67 sentence:T6-BTID-10098 head:T66 label:XCOMP:V-ADJ
R38 Dependency dependent:T74 sentence:T6-BTID-10098 head:T75 label:MOD_ATT:N-ADJ
R39 Dependency dependent:T72 sentence:T6-BTID-10098 head:T75 label:COORD_and:N-N
R40 Dependency dependent:T75 sentence:T6-BTID-10098 head:T70 label:OBJ:V-N
R41 Dependency dependent:T72 sentence:T6-BTID-10098 head:T70 label:OBJ:V-N
R42 Dependency dependent:T78 sentence:T6-BTID-10098 head:T79 label:MOD_ATT:N-ADJ
R43 Dependency dependent:T79 sentence:T6-BTID-10098 head:T77 label:OBJ:V-N
```

```
R44 Dependency dependent:T83 sentence:T6-BTID-10098 head:T85 label:MOD_ATT:N-N
R45 Dependency dependent:T87 sentence:T6-BTID-10098 head:T85 label:COMP_of:N-N
R46 Dependency dependent:T85 sentence:T6-BTID-10098 head:T82 label:OBJ:V-N
R47 Dependency dependent:T90 sentence:T6-BTID-10098 head:T82 label:COMP_through:V-N
R48 Dependency dependent:T81 sentence:T6-BTID-10098 head:T82 label:MOD:V-ADV
R49 Dependency dependent:T82 sentence:T6-BTID-10098 head:T77 label:COMP_before:V-V
R50 Dependency dependent:T77 sentence:T6-BTID-10098 head:T70 label:XMOD:V-V
R51 Dependency dependent:T65 sentence:T6-BTID-10098 head:T70 label:COORD_and:V-V
R52 Dependency dependent:T64 sentence:T6-BTID-10098 head:T66 label:SUBJ:V-N
R53 Dependency dependent:T64 sentence:T6-BTID-10098 head:T70 label:SUBJ:V-N
R54 Dependency dependent:T100 sentence:T7-BTID-10098 head:T101 label:MOD_ATT:N-N
R55 Dependency dependent:T99 sentence:T7-BTID-10098 head:T101 label:NUM:N-N
R56 Dependency dependent:T98 sentence:T7-BTID-10098 head:T101 label:MOD_ATT:N-N
R57 Dependency dependent:T105 sentence:T7-BTID-10098 head:T106 label:MOD_ATT:N-ADJ
R58 Dependency dependent:T106 sentence:T7-BTID-10098 head:T104 label:OBJ:V-N
R59 Dependency dependent:T110 sentence:T7-BTID-10098 head:T111 label:MOD_ATT:N-N
R60 Dependency dependent:T109 sentence:T7-BTID-10098 head:T111 label:MOD_ATT:N-ADJ
R61 Dependency dependent:T111 sentence:T7-BTID-10098 head:T108 label:OBJ:V-N
R62 Dependency dependent:T108 sentence:T7-BTID-10098 head:T104 label:COMP_by:V-V
R63 Dependency dependent:T103 sentence:T7-BTID-10098 head:T104 label:AUX:V-V
R64 Dependency dependent:T101 sentence:T7-BTID-10098 head:T104 label:SUBJ:V-N
R65 Dependency dependent:T101 sentence:T7-BTID-10098 head:T95 label:COMP_on:V-N
R66 Dependency dependent:T94 sentence:T7-BTID-10098 head:T95 label:AUX:V-V
R67 Dependency dependent:T93 sentence:T7-BTID-10098 head:T95 label:SUBJ:V-N
R68 Dependency dependent:T127 sentence:T10-BTID-10098 head:T125 label:COMP_of:N-N
R69 Dependency dependent:T125 sentence:T10-BTID-10098 head:T122 label:COMP_on:V-N
R70 Dependency dependent:T129 sentence:T10-BTID-10098 head:T122 label:COMP_by:V-N
R71 Dependency dependent:T120 sentence:T10-BTID-10098 head:T122 label:AUX:V-V
R72 Dependency dependent:T119 sentence:T10-BTID-10098 head:T122 label:SUBJ:V_PASS-N
R73 Dependency dependent:T131 sentence:T11-BTID-10098 head:T132 label:MOD_ATT:N-N
R74 Dependency dependent:T137 sentence:T11-BTID-10098 head:T138 label:MOD_ATT:N-N
R75 Dependency dependent:T138 sentence:T11-BTID-10098 head:T135 label:COMP_among:ADJ-N
R76 Dependency dependent:T143 sentence:T12-BTID-10098 head:T144 label:MOD_ATT:N-ADJ
R77 Dependency dependent:T146 sentence:T12-BTID-10098 head:T147 label:MOD_ATT:N-ADJ
R78 Dependency dependent:T147 sentence:T12-BTID-10098 head:T144 label:COMP_of:N-N
R79 Dependency dependent:T156 sentence:T12-BTID-10098 head:T154 label:COMP_of:N-N
R80 Dependency dependent:T152 sentence:T12-BTID-10098 head:T154 label:APPOS:N-N
R81 Dependency dependent:T158 sentence:T12-BTID-10098 head:T159 label:MOD_ATT:N-ADJ
R82 Dependency dependent:T152 sentence:T12-BTID-10098 head:T159 label:MOD_ATT:N-ADJ
R83 Dependency dependent:T163 sentence:T12-BTID-10098 head:T164 label:MOD_ATT:N-ADJ
R84 Dependency dependent:T166 sentence:T12-BTID-10098 head:T167 label:MOD_ATT:N-N
R85 Dependency dependent:T164 sentence:T12-BTID-10098 head:T167 label:COORD_and:N-N
R86 Dependency dependent:T167 sentence:T12-BTID-10098 head:T159 label:COMP_to:N-N
R87 Dependency dependent:T164 sentence:T12-BTID-10098 head:T159 label:COMP_to:N-N
R88 Dependency dependent:T159 sentence:T12-BTID-10098 head:T150 label:COMP_in:V-N
R89 Dependency dependent:T144 sentence:T12-BTID-10098 head:T150 label:SUBJ:V-N
R90 Dependency dependent:T144 sentence:T12-BTID-10098 head:T141 label:XCOMP:V-N
```

```
R91 Dependency dependent:T140 sentence:T12-BTID-10098 head:T141 label:SUBJ:V-N
R92 Dependency dependent:T170 sentence:T13-BTID-10098 head:T173 label:COMP_in:V-N
R93 Dependency dependent:T174 sentence:T13-BTID-10098 head:T175 label:MOD_ATT:N-ADJ
R94 Dependency dependent:T179 sentence:T13-BTID-10098 head:T180 label:MOD_ATT:N-ADJ
R95 Dependency dependent:T178 sentence:T13-BTID-10098 head:T180 label:MOD_ATT:N-ADJ
R96 Dependency dependent:T185 sentence:T13-BTID-10098 head:T186 label:MOD_ATT:N-ADJ
R97 Dependency dependent:T188 sentence:T13-BTID-10098 head:T186 label:COMP_of:N-N
R98 Dependency dependent:T186 sentence:T13-BTID-10098 head:T183 label:COMP_by:V-N
R99 Dependency dependent:T182 sentence:T13-BTID-10098 head:T183 label:AUX:V-V
R100 Dependency dependent:T180 sentence:T13-BTID-10098 head:T183 label:SUBJ:V_PASS-N
R101 Dependency dependent:T175 sentence:T13-BTID-10098 head:T180 label:APPOS:N-N
R102 Dependency dependent:T180 sentence:T13-BTID-10098 head:T173 label:OBJ:V-N
R103 Dependency dependent:T175 sentence:T13-BTID-10098 head:T173 label:OBJ:V-N
R104 Dependency dependent:T172 sentence:T13-BTID-10098 head:T173 label:SUBJ:V-N
R105 Dependency dependent:T191 sentence:T14-BTID-10098 head:T192 label:AUX:V-V
R106 Dependency dependent:T201 sentence:T14-BTID-10098 head:T202 label:MOD_ATT:N-N
R107 Dependency dependent:T200 sentence:T14-BTID-10098 head:T202 label:MOD_ATT:N-ADJ
R108 Dependency dependent:T197 sentence:T14-BTID-10098 head:T202 label:MOD_ATT:N-ADJ
R109 Dependency dependent:T195 sentence:T14-BTID-10098 head:T202 label:MOD_ATT:N-ADJ
R110 Dependency dependent:T202 sentence:T14-BTID-10098 head:T192 label:COMP_as:V_PASS-N
R111 Dependency dependent:T190 sentence:T14-BTID-10098 head:T192 label:SUBJ:V_PASS-N
R112 Dependency dependent:T212 sentence:T17-BTID-10098 head:T213 label:MOD:V-ADV
R113 Dependency dependent:T217 sentence:T17-BTID-10098 head:T215 label:COMP_by:V-N
R114 Dependency dependent:T214 sentence:T17-BTID-10098 head:T215 label:SUBJ:V_PASS-N
R115 Dependency dependent:T214 sentence:T17-BTID-10098 head:T215 label:SUBJ:V_PASS-N
R116 Dependency dependent:T214 sentence:T17-BTID-10098 head:T210 label:XCOMP:V-N
R117 Dependency dependent:T219 sentence:T18-BTID-10098 head:T220 label:MOD_ATT:N-N
R118 Dependency dependent:T222 sentence:T18-BTID-10098 head:T220 label:COMP_in:N-N
R119 Dependency dependent:T228 sentence:T19-BTID-10098 head:T230 label:COORD_or:ADJ-ADJ
R120 Dependency dependent:T230 sentence:T19-BTID-10098 head:T231 label:MOD_ATT:N-ADJ
R121 Dependency dependent:T228 sentence:T19-BTID-10098 head:T231 label:MOD_ATT:N-ADJ
R122 Dependency dependent:T227 sentence:T19-BTID-10098 head:T231 label:MOD_ATT:N-ADJ
R123 Dependency dependent:T236 sentence:T19-BTID-10098 head:T234 label:COMP_of:N-N
R124 Dependency dependent:T234 sentence:T19-BTID-10098 head:T232 label:OBJ:V-N
R125 Dependency dependent:T231 sentence:T19-BTID-10098 head:T232 label:SUBJ:V-N
R126 Dependency dependent:T231 sentence:T19-BTID-10098 head:T232 label:SUBJ:V-N
R127 Dependency dependent:T238 sentence:T19-BTID-10098 head:T237 label:NEG
R128 Dependency dependent:T239 sentence:T19-BTID-10098 head:T240 label:MOD:V-ADV
R129 Dependency dependent:T240 sentence:T19-BTID-10098 head:T237 label:XCOMP:V-V
R130 Dependency dependent:T231 sentence:T19-BTID-10098 head:T237 label:SUBJ:V-N
```

```
R131 Dependency dependent:T224 sentence:T19-BTID-10098 head:T237 label:MOD:V-ADV
R132 Dependency dependent:T243 sentence:T20-BTID-10098 head:T244 label:MOD_ATT:N-ADJ
R133 Dependency dependent:T244 sentence:T20-BTID-10098 head:T250 label:COMP_since:V-N
R134 Dependency dependent:T251 sentence:T20-BTID-10098 head:T250 label:OBJ:V_PASS-N
R135 Dependency dependent:T254 sentence:T20-BTID-10098 head:T255 label:MOD_ATT:N-ADJ
R136 Dependency dependent:T258 sentence:T20-BTID-10098 head:T257 label:OBJ:V-N
R137 Dependency dependent:T260 sentence:T20-BTID-10098 head:T257 label:COMP_in:V-N
R138 Dependency dependent:T266 sentence:T20-BTID-10098 head:T267 label:MOD_ATT:N-N
R139 Dependency dependent:T270 sentence:T20-BTID-10098 head:T267 label:COMP_during:N-N
R140 Dependency dependent:T267 sentence:T20-BTID-10098 head:T264 label:COMP_by:V-N
R141 Dependency dependent:T263 sentence:T20-BTID-10098 head:T264 label:AUX:V-V
R142 Dependency dependent:T262 sentence:T20-BTID-10098 head:T264 label:SUBJ:V_PASS-N
R143 Dependency dependent:T264 sentence:T20-BTID-10098 head:T257 label:COMP_until:V-V_PASS
R144 Dependency dependent:T255 sentence:T20-BTID-10098 head:T257 label:SUBJ:V-N
R145 Dependency dependent:T255 sentence:T20-BTID-10098 head:T257 label:SUBJ:V-N
R146 Dependency dependent:T255 sentence:T20-BTID-10098 head:T250 label:COMP_as:V-N
R147 Dependency dependent:T249 sentence:T20-BTID-10098 head:T250 label:AUX:V-V
R148 Dependency dependent:T248 sentence:T20-BTID-10098 head:T250 label:AUX:V-V
R149 Dependency dependent:T247 sentence:T20-BTID-10098 head:T250 label:SUBJ:V_PASS-N
R150 Dependency dependent:T273 sentence:T21-BTID-10098 head:T274 label:MOD_ATT:N-ADJ
R151 Dependency dependent:T277 sentence:T21-BTID-10098 head:T274 label:COMP_of:N-N
R152 Dependency dependent:T292 sentence:T21-BTID-10098 head:T287 label:COMP_through:V-N
R153 Dependency dependent:T294 sentence:T21-BTID-10098 head:T295 label:MOD_ATT:N-N
R154 Dependency dependent:T295 sentence:T21-BTID-10098 head:T287 label:COMP_of:V-N
R155 Dependency dependent:T300 sentence:T21-BTID-10098 head:T301 label:MOD_ATT:N-ADJ
R156 Dependency dependent:T301 sentence:T21-BTID-10098 head:T287 label:COMP_of:V-N
R157 Dependency dependent:T285 sentence:T21-BTID-10098 head:T287 label:COORD_or:V-N
R158 Dependency dependent:T287 sentence:T21-BTID-10098 head:T282 label:COMP_with:V_PASS-V
R159 Dependency dependent:T285 sentence:T21-BTID-10098 head:T282 label:COMP_with:V_PASS-N
R160 Dependency dependent:T281 sentence:T21-BTID-10098 head:T282 label:SUBJ:V_PASS-N
R161 Dependency dependent:T281 sentence:T21-BTID-10098 head:T282 label:SUBJ:V_PASS-N
R162 Dependency dependent:T281 sentence:T21-BTID-10098 head:T280 label:OBJ:V-N
R163 Dependency dependent:T280 sentence:T21-BTID-10098 head:T278 label:XCOMP:V-V
R164 Dependency dependent:T274 sentence:T21-BTID-10098 head:T278 label:SUBJ:V-N
R165 Dependency dependent:T274 sentence:T21-BTID-10098 head:T280 label:SUBJ:V-N
R166 Localization Bacterium:T310 Localization:T327
R167 Localization Bacterium:T310 Localization:T315
R168 Localization Bacterium:T320 Localization:T321
R169 Localization Bacterium:T310 Localization:T316
```

```
R170 Localization Bacterium:T312 Localization:T315
R171 Localization Bacterium:T309 Localization:T318
R172 Localization Bacterium:T312 Localization:T316
R173 Localization Bacterium:T309 Localization:T319
R174 Localization Bacterium:T310 Localization:T325
R175 Localization Bacterium:T311 Localization:T318
R176 Localization Bacterium:T311 Localization:T319
R177 Localization Bacterium:T309 Localization:T308
R178 Localization Bacterium:T311 Localization:T308
R179 Localization Bacterium:T309 Localization:T314
R180 Localization Bacterium:T311 Localization:T314
R181 Localization Bacterium:T309 Localization:T317
R182 Localization Bacterium:T311 Localization:T313
R183 Localization Bacterium:T311 Localization:T317
R184 Localization Bacterium:T309 Localization:T313
R185 Localization Bacterium:T312 Localization:T317
R186 Localization Bacterium:T312 Localization:T314
R187 Localization Bacterium:T312 Localization:T313
R188 Localization Bacterium:T310 Localization:T317
R189 Localization Bacterium:T310 Localization:T314
R190 Localization Bacterium:T310 Localization:T329
R191 Localization Bacterium:T305 Localization:T332
R192 Localization Bacterium:T310 Localization:T313
R193 Localization Bacterium:T312 Localization:T319
R194 Localization Bacterium:T312 Localization:T318
R195 Localization Bacterium:T312 Localization:T308
R196 Localization Bacterium:T311 Localization:T315
R197 Localization Bacterium:T311 Localization:T316
R198 Localization Bacterium:T309 Localization:T315
R199 Localization Bacterium:T309 Localization:T316
R200 Localization Bacterium:T310 Localization:T308
R201 Localization Bacterium:T310 Localization:T319
R202 Localization Bacterium:T310 Localization:T318
R203 Localization Bacterium:T310 Localization:T324
R167 PartOf Host:T308 Part:T315
R168 PartOf Host:T308 Part:T318
R169 PartOf Host:T308 Part:T321
R170 PartOf Host:T308 Part:T317
R171 PartOf Host:T308 Part:T316
R172 PartOf Host:T308 Part:T319
R168 Anaphora Anaphor:184-188 Ante:T307
R169 Anaphora Anaphor:251-259 Ante:T307
R170 Anaphora Anaphor:303-307 Ante:T307
R171 Anaphora Anaphor:345-357 Ante:T307
R172 Anaphora Anaphor:349-357 Ante:T307
R173 Anaphora Anaphor:380-382 Ante:T307
R174 Anaphora Anaphor:530-538 Ante:T307
R175 Anaphora Anaphor:722-735 Ante:T323
```

```
R176 Anaphora Anaphor:727-735 Ante:T323
R177 Anaphora Anaphor:839-841 Ante:T323
R178 Anaphora Anaphor:1013-1015 Ante:T323
R179 Anaphora Anaphor:1463-1474 Ante:T331
R180 Anaphora Anaphor:1468-1474 Ante:T331
R181 Anaphora Anaphor:1562-1564 Ante:T331
R182 Anaphora Anaphor:1581-1587 Ante:T331
```

# Appendix D

# The list of the articles in the *Arabidopsis Thaliana* corpus.

1. “MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes.” [Alvarez-Buylla et al., 2000]
2. “The MADS-Domain Protein AGAMOUS-Like 15 Accumulates in Embryonic Tissues with Diverse Origins” [Perry et al., 1999]
3. “LEAFY COTYLEDON1 represents a functionally specialized subunit of the CCAAT binding transcription factor” [Lee et al., 2003]
4. “Role of WRINKLED1 in the transcriptional regulation of glycolytic and fatty acid biosynthetic genes in Arabidopsis” [Baud et al., 2009]
5. “Control of expression and autoregulation of AGL15, a member of the MADS-box family” [Zhu and Perry, 2005]
6. “Indirect ABA-dependent Regulation of Seed Storage Protein Genes by FUSCA3 Transcription Factor in Arabidopsis” [Kagaya et al., 2005a]
7. “The Embryo MADS Domain Protein AGAMOUS-Like 15 Directly Regulates Expression of a Gene Encoding an Enzyme Involved in Gibberellin Metabolism” [Wang et al., 2004]
8. “Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis” [Braybrook et al., 2006]
9. “Repression of the LEAFY COTYLEDON 1/B3 Regulatory Network in Plant Embryo Development by VP1/ABSCISIC ACID INSENSITIVE 3-LIKE B3 Genes” [Suzuki et al., 2007]
10. “APETALA2 regulates the Stem Cell Niche in the Arabidopsis Shoot Meristem” [Würschum et al., 2006]
11. “AtGA3ox2, a Key Gene Responsible for Bioactive Gibberellin Biosynthesis, Is Regulated during Embryogenesis by LEAFY COTYLEDON2 and FUSCA3 in Arabidopsis” [Curaba et al., 2004]

12. “A Pivotal Role of the Basic Leucine Zipper Transcription Factor bZIP53 in the Regulation of Arabidopsis Seed Maturation Gene Expression Based on Heterodimerization and Protein Complex Formation” [Alonso et al., 2009]

13. “LEAFY COTYLEDON1 Controls Seed Storage Protein Genes through Its Regulation of FUSCA3 and ABSCISIC ACID INSENSITIVE3” [Kagaya et al., 2005b]

14. “The *turnip* Mutant of Arabidopsis Reveals That LEAFY COTYLEDON1 Expression Mediates the Effects of Auxin and Sugars to Promote Embryonic Cell Identity” [Casson and Lindsey, 2006]

15. “The FUS3 transcription factor functions through the epidermal regulator TTG1 during embryogenesis in Arabidopsis” [Tsuchiya et al., 2004]

16. “Effect of Regulated Overexpression of the MADS Domain Factor AGL15 on Flower Senescence and Fruit Maturation” [Fang and Fernandez, 2002]

17. “The ABSCISIC ACID INSENSITIVE 3 (ABI3) gene is modulated by farnesylation and is involved in auxin signaling and lateral root development in Arabidopsis” [Brady et al., 2003]

18. “Regulation and Function of the Arabidopsis ABA-insensitive4 Gene in Seed and Abscisic Acid Response Signaling Networks” [Söderman et al., 2000]

19. “The Transcription Factor FUSCA3 Controls Developmental Timing in Arabidopsis through the Hormones Gibberellin and Abscisic Acid” [Gazzarrini et al., 2004]

20. “The AIP2 E3 ligase acts as a novel negative regulator of ABA signaling by promoting ABI3 degradation” [Zhang et al., 2005]

21. “Synergistic Activation of Seed Storage Protein Gene Expression in Arabidopsis by ABI3 and Two bZIPs Related to OPAQUE2” [Lara et al., 2003]

22. “A subset of Arabidopsis AP2 transcription factors mediates cytokinin responses in concert with a two-component pathway” [Rashotte et al., 2006]

23. “Regulatory Networks in Seeds Integrating Developmental, Abscisic Acid, Sugar, and Light Signaling” [Brocard-Gifford et al., 2003]

24. “PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in Arabidopsis” [Ogas et al., 1999]

25. “A Network of Local and Redundant Gene Regulation Governs Arabidopsis Seed Maturation” [To et al., 2006]

26. “EMBRYONIC FACTOR 19 Encodes a Pentatricopeptide Repeat Protein that is Essential for the Initiation of Zygotic Embryogenesis in Arabidopsis” [Yu et al., 2012]

27. “Petunia Ap2-like genes and their role in flower and seed development” [Maes et al., 2001]

28. “LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development” [Stone et al., 2001]

29. “Arabidopsis LEAFY COTYLEDON1 Is Sufficient to Induce Embryo Development in Vegetative Cells” [Lotan et al., 1998]

30. “Mutual Regulation of Arabidopsis thaliana Ethylene-responsive Element Binding Protein and a Plant Floral Homeotic Gene, APETALA2” [Ogawa et al., 2007]

31. “Changes in gene expression in the leafy cotyledon1 (lec1) and fusca3 (fus3) mutants of Arabidopsis thaliana” [Vicient et al., 2000]

32. “Gene coexpression clusters and putative regulatory elements underlying seed storage reserve accumulation in Arabidopsis” [Peng and Weselake, 2011]

33. “A transcriptional repression motif in the MADS factor AGL15 is involved in recruitment of histone deacetylase complex components” [Hill et al., 2008]

34. “Control of seed mass by APETALA2” [Ohto et al., 2005]

35. “LEAFY COTYLEDON1-LIKE Defines a Class of Regulators Essential for Embryo Development” [Kwong et al., 2003]

36. “MUCILAGE-MODIFIED4 Encodes a Putative Pectin Biosynthetic Enzyme Developmentally Regulated by APETALA2, TRANSPARENT TESTA GLABRA1, and GLABRA2 in the Arabidopsis Seed Coat” [Western et al., 2004]

37. “WRI1 Is Required for Seed Germination and Seedling Establishment” [Cernac et al., 2006]

38. “Transcriptional Regulation of ABI3- and ABA-responsive Genes Including RD29B and RD29A in Seeds, Germinating Embryos, and Seedlings of Arabidopsis” [Nakashima et al., 2006]

39. “FUSCA3 encodes a protein with a conserved VP1/ABI3-like B3 domain which is of functional importance for the regulation of seed maturation in Arabidopsis thaliana” [Luerssen et al., 1998]

40. “MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis” [Nodine and Bartel, 2010]

41. “MicroRNAs Regulate the Timing of Embryo Maturation in Arabidopsis” [Willmann et al., 2011]

42. “The Embryo MADS Domain Factor AGL15 Acts Post embryonically: Inhibition of Perianth Senescence and Abscission via Constitutive Expression” [Fernandez et al., 2000]

43. “Physical interactions between ABA response loci of Arabidopsis” [Nakamura et al., 2001]

44. “An AP2-type transcription factor, WRINKLED1, of Arabidopsis thaliana binds to the AW-box sequence conserved among proximal upstream regions of genes involved in fatty acid synthesis.” [Maeo et al., 2009]

45. “The Arabidopsis SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE1 Protein Complex Includes BRASSINOSTEROID-INSENSITIVE1” [Karlova et al., 2006]

# Appendix E

# The *Arabidopsis Thaliana* Seed Development conceptual model in detail.

### 1. Entities

The definition of each entity type in this paragraph consists of: a short description, a list of short, characteristic examples and an example taken from the corpus, as well as the links to the sources and databases tied to this definition wherever applicable.

### Gene

A gene is a DNA sequence coding for a mRNA. In the *A. thaliana* literature, genes are always written in uppercase and italics.

Examples**:**

- *FUS3*
- *FUS3::GUS*
- *ABI3*
- *APETALA2*
- *At4g38130*
- In document #41 of the corpus: "We observed that in the wild type, the *FUS3* and *LEC2* transcriptional reporter genes are excluded from the embryo proper until the early heart stage"

### Gene Family

A family of genes mentioned by their common function, including coding for a same *protein family* or their common ancestor.

**Examples:**

- LEC genes
- AP2-like
- albumin genes
- In document #12 of the corpus: “The class of maturation genes (MAT) expressed during seed maturation typically includes seed storage protein (SSP) genes, such as albumin and cruciferin genes, which are induced in early or mid-maturation phase.”

**Sources:** TAIR[44]

### Promoter

A *Promoter* is an upstream region of a *Gene* that binds the polymerase for *Gene* transcription. It can be designated as a regulatory region of a *Gene*.

**Examples:**

- *BCCP2* promoter
- AGL15 regulatory regions
- 5’ flanking regions of LEC2-induced genes
- upstream region of *BCCP2*
- In document #44 of the corpus: “The WRI1 binding sites in the upstream regions of Pl-PK$\beta$1, KAS1, BCCP2, and SUS2 [...]”.

**Sources:** [Dérozier et al., 2011]

### RNA

*RNA* is a gene product.

**Examples:**

- *CLV3* mRNA
- LEC2-induced RNAs
- transcript of *FLC*
- In document #43 of the corpus: “ We initially isolated AGL15 as a low-abundance mRNA that preferentially accumulates in developing embryos.”

### Protein

A *Protein* is an RNA product. Proteins are always in plain letters and uppercase.

**Examples:**

- CLV3

[44] http://www.arabidopsis.org/browse/

- LEC1
- GLABRA3
- In document #44 of the corpus: "PLETHORA1 (PLT1) and PLT2 are required for stem cell specification and maintenance in the root meristem".

**Sources:** TAIR[44]

**Protein Family**

A family of proteins mentioned by their common biologic function or by their common ancestor.

**Examples:**

- MADS-domain proteins
- MYB
- bHLH protein
- bZIP transcription factor families
- seed storage proteins
- In document #29 of the corpus: "The LEC1 gene encodes a transcription factor homolog, the CCAAT box–binding factor HAP3 subunit."

**Sources:** [Jin et al., 2013, Davuluri et al., 2003]

**Protein Complex**

A *Protein Complex* is a group of *Proteins* that physically interact together.

**Examples:**

- Polycomb Group Protein (PCG)
- MYB-bHLH-WD40 (MBW)
- AFL (ABI3-FUS3-LEC)
- Core-binding factor (CBF)
- In document #28 of the corpus, "LEC1 shares extensive sequence similarity with the HAP3 subunit of CCAAT-binding transcription factor, implicating LEC1 as a transcriptional regulator".

**Sources:** TAIR[44]

**Protein Domain**

A *Protein Domain* is a protein sequence and structure that can evolve, function and exist independently.

**Examples:**

- B3 domain
- C-terminal region
- non-LEC1-type B domain
- basic helix-loop-helix (bHLH)
- In document #39 of the corpus, "Moreover, the C-terminal portion of FUS3 downstream of the B3 domain includes an acidic stretch and an amide sequence typically found in transcriptional activation domains."

**Sources:** TAIR[44]

### Hormone

A *Hormone* is a plant molecule that influences plant physiology and development.

**Examples:**

- auxin
- ethylene
- abscisic acid (ABA)
- In document #45 of the corpus, "In that work, it was clearly shown that the interaction between the BAK1 (SERK3) proteins and BRI1 is brassinolide-dependent"

**Sources:** TAIR[44]

### Regulatory Network

A regulatory network is defined as a set of *Products* and/or *DNA* that control the expression of a *gene*, a *pathway.* It can also be used as a term to describe processes and functionality involving several *genes.* In all of these cases, a *regulatory network* corresponds to a regulatory function. This includes signalling pathways.

**Examples:**

- the sensitivity or response to a factor *(e.g. sensitivity to abscisic acid or sugar responses)*
- the acquisition of properties (e.g. *acquisition of desiccation tolerance*).
- Processes and functionality involving several genes (e.g. *meristem function*, *control of seed size*, *maturation processes*)
- The genes that regulate *Pathways*, *Development Phase, Regulatory Network,* like the genes that regulate *female reproductive tract development.*
- The genes involved in/related to *Development Phase, Regulatory Network,* like the genes involved in *embryo formation.*
- In document #40 of the corpus, "These embryos with defects were presumably homozygous for dcl1-5-null alleles, and their defects included abnormal hypophysis cell divisions as well

as the previously unreported loss of periclinal subprotoderm cell divisions in the embryo proper."

**Sources:** TAIR[44]

### Pathway

*Pathway* means here metabolic pathway, for instance synthesis or degradation. A *pathway* represents a group of *genes* or corresponding *products* that are involved in a same metabolic, physiological or developmental *pathway.*

**Examples:**

- metabolic pathways
- fatty acid pathways
- flavonoid pathway
- In document #4 of the corpus, "The WRI1 transcription factor, [...], was proposed to trigger the transcription of the set of genes involved in the conversion of sucrose into fatty acids."

**Sources:** Pathways database[45] from Carnegie Science.

### Genotype

This types covers both genotypes and species.

A *genotype* is given part or the whole genetic information (genetic composition) expressed by an organism genome. In this case, for Arabidopsis thaliana.

A species is defined in reference to the biological nomenclature.

**Examples:**

- serk2 null mutant
- variety
- cultivar
- cyanobacteria
- maize
- overexpression of WRI1
- fus3
- In document #4 of the corpus, "This hypothesis was based on the analysis of the expression profiles of putative target genes of WRI1 characterized in various *wri1* mutant backgrounds, and in tissues overexpressing *WRI1* ectopically."

**Sources:** TAIR[44]

[45] https://ftp.dpb.carnegiescience.edu//Pathways/

### Tissue

This type groups cell, tissue and organ.

A *Tissue* is an ensemble of cells, not necessarily identical, that together carry out a specific function. Organs are then formed by the functional grouping together of multiple tissues. The *Tissue* type includes organs, as well as entities on the intra-cellular level, such as the "nuclei of the embryos" for example.

**Examples:**

- vegetative tissues
- seedling
- endosperm
- mature seed
- cytoplasm
- In document #29 of the corpus, "*LEC1* RNA accumulates only during seed development shoot or root apical meristems in embryo cell types and in endosperm tissue"

**Sources:** SeedGeneNetwork[46],[Winter et al., 2007], eFP browser[47]

### Development Phase

A growth stage. This includes identity (cotyledon identity), dormancy (bud dormancy), development (cotyledon development) and growth (etiolated growth).

**Examples:**

- number of days after fertilization (daf)
- flowering
- Maturation
- In document #40 of the corpus, "*WUSCHEL-related HOMEOBOX2 (WOX2)* transcripts are localized in the apical cell lineage of wild-type preglobular embryos."

**Sources:** TAIR[44]

### Environmental Factor

Environmental or experimental conditions. This means any factor, abiotic or biotic, that influences living organisms (e.g. temperature, light, "in vitro").

**Examples:**

- temperature
- shade

[46] http://seedgenenetwork.net/

[47] http://bar.utoronto.ca/efp2/Arabidopsis/Arabidopsis_eFPBrowser2.html

- carbon
- sugar
- biotic stress
- Pathogen
- In document #45 of the corpus, “The interaction between SERK1 and BRI1 was confirmed by a genetic experiment and fluorescence lifetime imaging microscopy (FLIM) to determine Förster resonance energy transfer (FRET) between fluorescently tagged receptors”.

**Sources:** TAIR[44]

***Note*:** The following list defines categories of types that have been defined for expressing the constraints on relation argument types in a concise way.

### DNA

A type grouping *Gene*, *Gene Family*, *Box* and *Promoter.*

### Amino Acid Sequence

A type grouping *Protein*, *Protein Family*, *Protein Complex* and *Protein Domain.*

### Gene Product

A type grouping *RNA* and *Amino Acid Sequence.*

### Functional Molecule

A type grouping *Gene Product* and *Hormone.*

### Molecule

A type grouping together *DNA*, and *Functional Molecule.*

### Dynamic Process

A type grouping *Regulatory Network* and *Pathway.*

### Biological Context

This group contains all biological entities which play the role of a factor in events. This means *Genotype*, *Tissue* and *Development Phase.*

### Context

This group contains all possible factors, ie *Biological Context* and *Environmental Factor.*

## 2. Relations

For clarity, in the definitions below, argument names are written in **bold** letters, whereas entity or relation types are written in *italic.*

### Time and Localization

#### *Presence In Genotype*

A **Molecule** or **Element** is present in a given **Genotype**.

**Arguments**:

- **Genotype**: [*Genotype*], and
  - **Molecule**:[*Molecule*], or
  - **Element**: [*Biological Context*]

**Examples:**

- (Document #4) Proteins containing homologues of the AP2 domain [*Protein Domain*] have been identified in cyanobacteria [*Genotype*]
- (Document #33) Yeast [*Genotype*] has only one HDAC [*Protein Family*]

#### *Occurrence In Genotype*

A **Process** occurs in a given **Genotype**.

**Arguments**:

- **Process**: [*Dynamic Process*]
- **Genotype**: [*Genotype*]

**Examples:**

- (Document #33) It is becoming increasingly apparent that autoregulatory loops are a common phenomenon in the regulation of MADS-box genes [*Regulatory Network*] in plants [*Genotype*]

#### Presence At Stage

A *Functional Molecule* is present during a given *Development phase.*

**Arguments**:

- **Functional Molecule**: [*Functional Molecule*]
- **Development** : [*Development Phase*]

**Examples:**

- (Document #28) We showed that *LEC2* RNA [*RNA*] accumulates primarily during seed development [*Development Phase*]

#### ***Occurrence During***

A **Process** occurs during a given *Development Phase.*

**Arguments**:

- **Process**: [*Dynamic Process*]
- **Development**:[*Development Phase*]

**Examples:**

- (Document #29) Higher plant embryogenesis [*Development Phase*] is divided conceptually into two distinct phases: early morphogenetic processes [*Regulatory Network*] [...]

#### **Localization**

A *Functional Molecule* or *Dynamic Process* is found in a *Tissue***.**

**Arguments:**

- **Target Tissue**: [*Tissue*], and
  - **Functional Molecule**: [*Functional Molecule*], or
  - **Process**: [*Dynamic Process*]

**Examples:**

- FUS3 mRNA [*RNA*] accumulates in seed [*Tissue*]
- AGL15 [*Protein*] was initially present in the cytoplasm of cells [*Tissue*]

### **Function**

#### **Involvement In Process**

A *Molecule* is involved in a *Dynamic Process.*

**Arguments:**

- **Participant:** [*Molecule*]
- **Process:** [*Dynamic Process*]

**Examples:**

- (Document #36) A complex process of differentiation [*Regulatory Network*] that includes the biosynthesis and secretion of pectinaceous mucilage [*Pathway*]
  - biosynthesis of pectinaceous mucilage [*Pathway*] Involvement In Process process of differentiation [*Regulatory Network*]
  - secretion of pectinaceous mucilage [*Pathway*] Involvement In Process process of differentiation [*Regulatory Network*]
- (Document #36) *MUM4* [*Gene*] encodes an enzyme involved in RGI biosynthesis[*Pathway*]

### Transcription Or Translation

*DNA* entities encode for *RNA* (Transcription) or *RNA* entities encode *Proteins* (Translation). Often, reference is made to the *Gene* encoding the protein, without mention of the *RNA*.

**Arguments:**

- **Source:** [*DNA* | *RNA*]
- **Product:** [*Gene Product*]

**Examples:**

- (Document #9) Three VP1/ABI3-LIKE (VAL) [*Gene Family*] genes encode B3 proteins [*Protein Family*]

### Functional Equivalence

A *Molecule*, *Dynamic Process* or *Context* compared to another similar *Molecule*, *Dynamic Process* or Context. This type is used to link similar products in different species, such as homolog or ortholog *Proteins.*

**Arguments:**

- **Element1**: [Any entity type]
- **Element2**: [Any entity type]

**Examples :**

- *WER* [*Gene*] and *GL1* [*Gene*] encode functionally equivalent proteins
- (Document #43) Rice homologs of ABI3 and ABI5 (OSVP1 and TRAB1, respectively)
    - ABI3 [*Protein*] Functional Equivalence OSVP1 [*Protein*]
    - ABI5 [*Protein*] Functional Equivalence TRAB1 [*Protein*]

### Regulation

*Regulation* relation types are used when there is a *Genotype* involved. The **Agent** of this type of relation is always the *Genotype*, even when the *Gene* involved in the *Genotype* is specified, if no information on the direct role of the agent is given. If direct information on the **Agent** is given, genotype is indicated in the optional argument **Organism Genotype** [*Genotype*]**.**

**Examples :**

- (Document #9) RNA interference of *L1L* function [*Genotype*] has been shown to cause embryo arrest [*Regulatory Network*]
- (Document #9) Ectopic expression of LEC1 [*Genotype*] is sufficient to induce embryo formation [*Regulatory Network*]
- (Document #9) *val1 val2* double-mutant [*Genotype*] seedlings form no leaves [*Tissue*]

### Regulation Of Accumulation

An **Agent** regulates the accumulation of a *Molecule.*

**Arguments:**

- **Agent**: [Any entity type]
- **Functional Molecule:** [*Functional Molecule*]

**Examples:**

- (Document #33) AGL15-VP16 [*Protein Complex*] induces accumulation of AGL18 transcript [*RNA*]
- (Document #2) Finally, the results of our studies of somatic embryogenesis [*Development Phase*] indicate that AGL15 [*Protein*] accumulates even when embryos [*Tissue*] arise de novo from cells in other phases of the life cycle.
  - *Regulation Of Accumulation* **: Agent=** somatic embryogenesis [*Development Phase*] + **Functional Molecule=** AGL15 [*Protein*] + **Tissue** =embryos [*Tissue*]

### Regulation Of Expression

An **Agent** regulates the expression of a *DNA* entity.

**Arguments:**

- **Agent:** [Any entity type]
- **DNA:** [*DNA*]

**Examples:**

- (Document #4) WRI1 [*Protein*] is a limiting factor of lipogenic gene [*Gene Family*] expression in seeds [*Tissue*], directly induces the transcriptional activation of these genes at the onset of the maturation phase [*Development Phase*].
  - Regulation Of Expression: **Agent=** WRI1 [*Protein*] + **DNA=** lipogenic gene [*Gene Family*] + **Tissue=** seeds [*Tissue*] + **Development Stage=**maturation phase [*Development Phase*]
- (Document #29) Expression of the *LEC1* [*Gene*] gene in vegetative cells [*Tissue*]
- (Document #4) We suggest that VAL [*Gene Family*] targets Sph/RY [*Box*] -containing genes

### Regulation Of Development Phase

An **Agent** regulates the activity of a *Development Phase.*

**Arguments:**

- **Agent:** [Any entity type]
- **Development:** [*Development Phase]*

**Examples:**

- Biologically active GAs [*Hormone*] have been suggested to have a role in embryogenesis [*Development Phase*]
- Genetic analysis shows that termination of the primary shoot meristem [*Development Phase*] in l28 mutants [*Genotype*] requires an active CLV signaling pathway [*Pathway*]
  - Regulation Of Development Phase: **Agent=** CLV signaling pathway [*Pathway*] + **Development=** termination of the primary shoot meristem [*Development Phase*] + **Organism Genotype=** l28 mutants [*Genotype*]

### Regulation Of Molecule Activity

An **Agent** regulates the activity of a **Molecule**.

**Arguments:**

- **Agent:** [Any entity type]
- **Molecule:** [*Amino Acid Sequence* | *Hormone*]

**Examples:**

- (Document #45) p97/VCP [*Protein Complex*] can be phosphorylated by the JAK-2 [*Protein*] kinase

### Regulation Of Process

An **Agent** regulates the activity of a *Dynamic Process.*

**Arguments:**

- **Agent:** [Any entity type]
- **Process:** [Dynamic Process]

**Examples:**

- (Document #10) The stem cells in turn signal back via the CLV3 [*Protein*] peptide to restrict the size of the OC [*Regulatory Network*]
- (Document #4) WRI1 [*Protein*] directly enhances the expression of genes involved in glycolysis [*Pathway*]

### Regulation Of Tissue Development

An **Agent** regulates the development of a *Tissue.*

**Arguments:**

- **Agent:** [Any entity type]
- **Target Tissue:** [*Tissue*]

**Examples:**

- (Document #2) The results of our study of organs [*Tissue*] produced during precocious germination [*Regulatory Network*]

### Composition and Membership

#### Primary Structure Composition

A specific sequence of nucleotide is found in a molecule of **DNA.**

**Arguments:**

- **DNA Part***:* [*Box* / *Promoter*]
- **DNA:** [*DNA*]

**Examples:**

- (Document #4) We show that mutations in the AACCCA [*Box*] element of the *BCCP2* promoter [*Promoter*]
- (Document #44) The WRI1 binding sites in the upstream region of Pl-PK$\beta$1 [*Promoter*] contained the conserved AW-box [*Box*]

#### Protein Complex Composition

An *Amino Acid Sequence* is found in a *Protein Complex***.**

**Arguments:**

- **Amino Acid Sequence:** [*Amino Acid Sequence*]
- **Protein Complex:** [*Protein Complex*]

**Examples:**

- (Document #45) The identification of two members of the BR signaling pathway, the main BR receptor BRI1 [*Protein*] and its coreceptor BAK1 [*Protein*] (SERK3 [*Protein*]) as components of the SERK1 complex [*Protein Complex*].

#### Protein Domain Composition

A specific *Protein Domain* is found in an *Amino Acid Sequence.* It can be used to link products that are part of a factor.

**Arguments:**

- **Domain:** [*Protein Domain*]
- **Product:** [*DNA Product*]

**Examples:**

- (Document #4) Proteins of the RAV family [*Protein Family*] contain one AP2 domain [*Protein Domain*]

### Family Membership

A *DNA*, or *Gene Product* belongs to another *DNA*, or *Gene Product.* This relation is to be used between entities of the same nature, to denote members of a set (e.g. [*Gene*] belonging to [*Gene Family*], [*Protein*] to a [*Protein Family*], sub-families to families, etc.).

**Arguments:**

- **Element:** [*DNA* | *Gene Product*]
- **Family:** [*Gene Family* / *RNA* | *Protein Family*]

**Examples:**

- (Document #30) AP2 [*Protein*] belongs to AP2/EREBP [*Protein Family*] family
- (Document #42) *AGL15* [*Gene*] from the large group of floral MADS box genes [*Gene Family*]
- (Document #12) bZIP10 [*Protein*] and bZIP25 [*Protein*], which have been classified into group C [*Protein Family*]

### Sequence Identity

A *Molecule*, *Dynamic Process* or *Context* compared to another similar *Molecule*, *Dynamic Process* or *Context.* This type of relation is used for linking identical products, as well as synonyms, full form and abbreviation.

**Arguments:**

- **Element1**: [Any type of entity]
- **Element2**: [Any type of entity]

**Examples :**

- The SUN6 gene [*Gene*] is identical to the previously described ABI4 gene [*Gene*]
- (Document #4) LEAFY COTYLEDON2 [*Protein*] (LEC2)[*Protein*]

### Interaction

### Binding

A *Functional Molecule* physically binds to a *Molecule.* In most cases, a *Protein* binds to a *Promoter* or a *Gene.* An interaction between two protein is specifically performed "*in vitro*" or in "yeast two-hybrid" is annotated as a Binding relation. Exceptionally, in the specific case of a homodimeric interaction, both arguments of the Binding relation are the same entity:

**Arguments:**

**Functional Molecule:** [*Functional Molecule*]

**Molecule:** [*Molecule*]

**Examples:**

- (Document #4) Interaction between WRI1 [*Protein*] and the *BCCP2* promoter [*Promoter*], both *in vitro* and in yeast
- (Document #33) TT2 [*Protein*] interacts with TT8-TTG1 [*Protein Complex*] in yeast two-hybrid studies.
- (Document #10) CLV3 [*Protein*] acts as an extracellular ligand of the CLV1 receptor kinase complex [*Protein Complex*]
- FUS3 [*Protein*] interacts with LEC2 [*Protein*] *in vitro*
- (Document #45) BRI1 [*Protein*] can also form homodimers in the plasma membrane [*Tissue*]
  - In this case: Binding **Functional Molecule=** BRI1 [*Protein*] + **Molecule=** BRI1 [*Protein*] + **Tissue=**plasma membrane [*Tissue*]

### Interaction

A *Molecule* interacts with another *Molecule.* This type is used between DNA-DNA, in the case of indirect (non physical) interaction, and in any case of interaction where a more specific relation type cannot be used.

**Arguments:**

**Agent:** [*DNA* | *Amino Acid Sequence*]

**Target:** [*DNA* | *Amino Acid Sequence*]

**Examples:**

- (Document #33) *SAP18* [*Gene*] alone repressed *LEA* and *CBF2*, possibly through interaction with *AGL15* [*Gene*]
- (Document #33) AGL15 [*Protein*] interacts with members of the SIN3 histone deacetylase (HDAC) complex [*Protein Complex*] (not explicitly a physical bind)

### Secondary Arguments

These optional secondary arguments are used to describe complex n-ary events and could be serve the role of conditions or restraints. There are six types of secondary arguments, five for entities and one for events. Only one entity per role is possible for the entity secondary arguments, whereas multiple secondary events can be linked. All event types apart from *Presence in Genotype* accepttake secondary arguments.

1. **Tissue**: [*Tissue*]
2. **Development Stage**: [*Development Phase*]
3. **Organism Genotype**: [*Genotype*]
4. **Environmental Factor**: [*Environmental Factor*]
5. **Hormone**: [*Hormone*]

6. **Prerequisite Event:** [*Primary Structure Composition | Interaction | Localization | Protein Domain Composition*]

**Examples:**

- Entity argument: "A [*Protein*] accumulates in B [*Tissue*] when there exists C [*Hormone*]."
  - R1: *Localization* (**Functional Molecule**: A, **Target Tissue**: B, **Hormone** : C).
- Two entity arguments: "A [*Protein*] activates B [*Gene*] in the flower [*Tissue*], if C [*Hormone*] increases."
  - R1: *Regulation Of Expression* (**Agent**: A, **DNA**: B , **Tissue**: flower , **Hormone**: C).
- Conjunction → two entity arguments: "A [*Protein*] accumulates in B [*Tissue*] when there exists C [*Hormone*] *and* D [*Environmental Factor*]."
  - R1: *Localization* (**Functional Molecule**: A, **Target Tissue**: B, **Hormone**: C, **Environmental Factor**: D).
- Disjunction → two unlinked relations: "A [*Protein*] accumulates in B [*Tissue*] when there exists C [*Hormone*] *or* D [*Environmental Factor*]."
  - R1: *Localization* (**Functional Molecule**: A, **Target Tissue**: B, **Hormone**: C).
  - R2 : *Localization* (**Functional Molecule**: A, **Target Tissue**: B, **Environmental Factor**: D).
- Relation argument: "A [*Protein*] binds to B [*Protein complex*] if C [*Hormone*] is found in D [*Tissue*]."
  - R1: *Localization* (**Functional Molecule**: B, **Target Tissue** D [*Tissue*])
  - R2: *Binding* **(Agent**: A, **Target**: B, **Prerequisite Event**: R1)

# Appendix F

# The *Arabidopsis Thaliana* Seed Development annotation model in detail.

Entities are the same as above (Appendix E).

**Relations**

**Interaction**

**Binds To**

A *Product* physically binds to *DNA* or another *Product*. In most cases, a *Protein* binds to a *Promoter* or a *Gene*.

**Arguments:**

- **Agent**: [*Product*]
- **Target**: [*DNA* | *Product*]

**Parameters:**

- Binding (default value)
- Increase
- Decrease

**Examples:**

- FUS3 [*Protein*] binds pAt2S3 [*Promoter*]
- TT2 [*Protein*] interacts with TT8-TTG1 [*Protein Complex*] in yeast two-hybrid studies.
- WRI1 [*Protein*] was able to interact with the BCCP2 [*Promoter*]

**Interacts With**

A *DNA*, *Product* or *Factor* interacts with another *DNA*, *Product*, or *Factor*, directly or indirectly. It is used whenever the more specialized *BindsTo* (for direct physical interactions), *Regulates Activity Of* and *Regulates Expression Of* are not appropriate.

**Arguments:**

- **Agent**: [*DNA* | *Product* | *Factor*]
- **Target**: [*DNA* | *Product* | *Factor*]

**Parameters:**

- Interaction (default value)
- Increase
- Decrease

**Examples:**

- RY box [*Box]* interacts with G box [*Box]* (Genetic interaction*)*
- FUS3 [*Protein]* interacts with LEC2 [*Protein]* (not explicitly a physical bind)

**Similarity**

**Encodes**

A *DNA* or *RNA* entity encodes a *Product* such as a *Protein.*

**Arguments**:

- **Agent**: [*DNA* | *Product*]
- **Target**: [*Product*]

**Examples:**

- Three VP1/ABI3-LIKE (VAL) [*Gene Family*] genes encode B3 proteins [*Protein Family*]
- LEC2 RNA [RNA] encodes a regulator of reserve accumulation, EEL [*Protein*].

**Belongs To**

A *DNA, Product* or *Factor* belongs to another *DNA*, *Product* or *Factor.* This relation is to be used between entities of the same nature, to denote members of a set (e.g. genes belonging to gene family, proteins to a protein family, subfamilies to families, etc.).

**Arguments**:

- **Agent**: [*DNA* | *Product* | *Factor*]
- **Target**: [*DNA* | *Product* | *Factor*]

**Examples:**

- FUSCA3 [*Gene*] belongs to the B3 [*Gene Family*]
- bZIP10 [*Protein*] and bZIP25 [*Protein*], which have been classified into group C [*Protein Family*]

### Comparison

A *DNA*, *Product* or *Factor* compared to another *DNA*, *Product* or *Factor*

This relation type is also used to link abbreviations to their full form. It has three specifications a) equivalent in function, b) identical in sequence and redundant which denotes equivalence in time, localization and function at the same time.

**Arguments:**

- **Agent**: [*DNA* | *Product* | *Factor*]
- **Target**: [*DNA* | *Product* | *Factor*]

**Parameters:**

- equivalent
- identical
- redundant

**Examples**:

- WER [*Gene]* and GL1 [*Gene*] encode functionally equivalent proteins
- The SUN6 gene [*Gene*] is identical to the previously described ABI4 gene [*Gene*]
- Arabidopsis myrosinases TGG1 [*Gene*] and TGG2 [*Gene*] have redundant function in glucosinolate breakdown and insect defense.
- LEAFY COTYLEDON2 [*Protein*] (LEC2)[*Protein*]

### Localization

### Is Found In or During

A *Product* accumulates or is found in a given *Factor* or during a *Development Phase.* In reality this type can be directly split into two distinct types, *Is Found In* and *Is Found During*, depending on the type of the second argument, but these two types were merged in order to reduce the size of the annotation model.

**Arguments:**

- **Agent**: [*Product*]
- **Target**: [*Factor*]

**Parameters:**

- Presence (default value)
- Increase
- Decrease

**Examples:**

- Proteins of the RAV family [*Protein Family*] contain one AP2 domain [*Protein Domain*]
- Proteins containing homologues of the AP2 domain [*Protein Domain*] have been identified in cyanobacteria [*Genotype*]
- AGL15 [*Protein*] was initially present in the cytoplasm of cells [*Tissue*]

### Regulation

#### Regulates Activity Of

A *DNA*, *Product* or *Factor* regulates the activity of a *Product* or a *Factor*. This relation type is used whenever it is not possible to use the more specific types *Regulates Expression Of* and *Regulates Accumulation Of*.

**Arguments:**

- **Agent**: [*DNA* | *Product* | *Factor*]
- **Target**: [*Product* | *Factor*]

**Parameters:**

- Involvement (default value)
- Activation
- Inhibition
- Requirement

**Examples:**

- ABA [*Hormone*] activates At2S2 [*Gene*] gene expression
- Biologically active GAs [*Hormone*] have been suggested to have a role in embryogenesis [*Development Phase*]
- Fungi [*Environmental Factor*] activate the TT regulatory network [*Regulatory Network*]

#### Regulates Expression Of

A *DNA*, *Product* or *Factor* regulates the activity of a *DNA* entity.

**Arguments:**

- **Agent**: [*DNA* | *Product* | *Factor*]

- **Target**: [*DNA*]

**Parameters:**

- Involvement (default value)
- Activation
- Inhibition
- Requirement

**Examples:**

- WRI1 [*Protein*] regulates the activity of PKp-$\beta$1 promoter [*Promoter*]
- FUS3-mediated [*Protein*] induction of CRC [*Gene*]
- WRI1 [*Protein*] is a limiting factor of lipogenic gene [*Gene Family*] expression in seeds
- LEC2 [*Gene*] expression is normally limited primarily to seed development [*Development Phase*]

### Regulates Accumulation Of

A *DNA*, *Product* or *Factor* regulates the accumulation of a *Product*, and more specifically of a *Protein*, *RNA*, or a *Hormone*.

**Arguments:**

- **Agent**: [*DNA* | *Product* | *Factor*]
- **Target**: [*Product*]

**Parameters:**

- Involvement (default value)
- Activation
- Inhibition
- Requirement

**Examples:**

- FUS3 [*Protein*] activates the accumulation of At2S3 [*Protein*]
- Induction of LEC2 activity in seedlings [*Genotype*] causes rapid accumulation of RNAs normally present primarily during the maturation phase [*RNA*]

### N-ary Event

**Condition**

Condition is the only type of relation in this corpus which can take another relation as an argument. It is used to emulate multi-argument events and it is comparable to the mechanism of secondary arguments in the conceptual model.

**Arguments:**

- **Relation**: [Any relation type]
- **Constraint**: [See below]

The second argument, **Constraint**, can be any of the following types:

- Entities:
  - *Hormone*
  - *Tissue*
  - *Development Phase*
  - *Environmental Factor*
  - *Genotype*
- Relations:
  - *Is Found In or During*
  - *Binds To*
  - *Interacts With*

**Examples**:

- Relation - Entity: "A [*Protein*] accumulates in B [*Tissue*] when there exists C [*Hormone*]."
  - R1: *Is Found In or During*(**Agent**: A, **Target**: B)
  - R2: *Condition*(**Relation**: R1, **Constraint**: C)
- Relation - Relation: "A [*Protein*] activates B [*Gene*] in the flower [*Tissue*], if C [*Hormone*] increases."
  - R1: *Regulates Activity Of*(**Agent**: A, **Target**: B)
  - R2: *Is Found In or During*(**Agent**: B, **Target**: flower)
  - R3: *Is Found In or During*(**Agent**: A, **Target**: flower)
  - R4: *Condition*(**Relation**: R1, **Constraint**: R3)
- Double Condition: "A [*Protein*] accumulates in B [*Tissue*] when there exists C [*Hormone*] and D [*Environmental Factor*]."
  - R1: *Regulates Activity Of*(**Agent**: A, **Target**: B)
  - R2: *Condition*(**Relation**: R1, **Constraint**: C)
  - R3: *Condition*(**Relation**: R1, **Constraint**: D)

# Bibliography

[Abulaish and Dey, 2007] Abulaish, M. and Dey, L. (2007). Biological relation extraction and query answering from MEDLINE abstracts using ontology-based text mining. *Data Knowl. Eng.*, 61(2):228–262.

[Ackoff, 1989] Ackoff, R. L. (1989). From data to wisdom. *J. Appl. Syst. Anal.*, 16:3–9.

[Agichtein and Gravano, 2000] Agichtein, E. and Gravano, L. (2000). Extracting relations from large Plain-Text collections. *Proceedings of International Conference on Digital Libraries*, I(58):85–94.

[Airola et al., 2008] Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., and Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9 Suppl 11:S2.

[Aldana and Cluzel, 2003] Aldana, M. and Cluzel, P. (2003). A natural class of robust networks. *Proc. Natl. Acad. Sci. U. S. A.*, 100(15):8710–8714.

[Alex et al., 2010] Alex, B., Grover, C., Shen, R., and Kabadjov, M. (2010). Agile corpus annotation in practice: An overview of manual and automatic annotation of CVs. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 29–37, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Alonso et al., 2009] Alonso, R., Oñate Sánchez, L., Weltmeier, F., Ehlert, A., Diaz, I., Dietrich, K., Vicente-Carbajosa, J., and Dröge-Laser, W. (2009). A pivotal role of the basic leucine zipper transcription factor bZIP53 in the regulation of arabidopsis seed maturation gene expression based on heterodimerization and protein complex formation. *Plant Cell*, 21(6):1747–1761.

[Alvarez-Buylla et al., 2007] Alvarez-Buylla, E. R., Benítez, M., Dávila, E. B., Chaos, A., Espinosa-Soto, C., and Padilla-Longoria, P. (2007). Gene regulatory network models for plant development. *Curr. Opin. Plant Biol.*, 10(1):83–91.

[Alvarez-Buylla et al., 2000] Alvarez-Buylla, E. R., Liljegren, S. J., Pelaz, S., Gold, S. E., Burgeff, C., Ditta, G. S., Vergara-Silva, F., and Yanofsky, M. F. (2000). MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. *Plant J.*, 24(4):457–466.

[Ananiadou et al., 2006] Ananiadou, S., Kell, D. B., and Tsujii, J.-I. (2006). Text mining and its potential applications in systems biology. *Trends Biotechnol.*, 24(12):571–579.

[Ananiadou et al., 2010] Ananiadou, S., Pyysalo, S., Tsujii, J., and Kell, D. B. (2010). Event extraction for systems biology by text mining the literature. *Trends Biotechnol.*, 28(7):381–390.

[Ananiadou et al., 2004] Ananiadou, S., Sophia, A., Carol, F., and Jun'ichi, T. (2004). Introduction: named entity recognition in biomedicine. *J. Biomed. Inform.*, 37(6):393–395.

[Arighi et al., 2013] Arighi, C. N., Carterette, B., Cohen, K. B., Krallinger, M., Wilbur, W. J., Fey, P., Dodson, R., Cooper, L., Van Slyke, C. E., Dahdul, W., Mabee, P., Li, D., Harris, B., Gillespie, M., Jimenez, S., Roberts, P., Matthews, L., Becker, K., Drabkin, H., Bello, S., Licata, L., Chatr-Aryamontri, A., Schaeffer, M. L., Park, J., Haendel, M., Van Auken, K., Li, Y., Chan, J., Muller, H.-M., Cui, H., Balhoff, J. P., Chi-Yang Wu, J., Lu, Z., Wei, C.-H., Tudor, C. O., Raja, K., Subramani, S., Natarajan, J., Cejuela, J. M., Dubey, P., and Wu, C. (2013). An overview of the BioCreative 2012 workshop track III: interactive text mining task. *Database*, 2013:bas056.

[Aronson and Lang, 2010] Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*, 17(3):229–236.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Michael Cherry, J., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1):25–29.

[Aubin and Hamon, 2006] Aubin, S. and Hamon, T. (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, pages 380–387. Springer.

[Ba and Bossy, 2016] Ba, M. and Bossy, R. (2016). Interoperability of corpus processing workflow engines: the case of AlvisNLP/ML in OpenMinTeD. In Eckart de Castilho, R., Ananiadou, S., Margoni, T., Peters, W., and Piperidis, S., editors, *Proceedings of the LREC 2016 Workshop "Cross-Platform Text Mining and Natural Language Processing Interoperability"*.

[Bada et al., 2012] Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K., Verspoor, K., Blake, J. A., and Hunter, L. E. (2012). Concept annotation in the CRAFT corpus.

[Balbi and Devoto, 2007] Balbi, V. and Devoto, A. (2007). Jasmonate signalling network in arabidopsis thaliana: crucial regulatory nodes and new physiological scenarios. *New Phytol.*, 177(2):301–318.

[Banerjee and Pedersen, 2002] Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using WordNet. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 136–145. Springer Berlin Heidelberg.

[Banko et al., 2008] Banko, M., Etzioni, O., and Center, T. (2008). The tradeoffs between open and traditional relation extraction. *Proceedings of ACL-08: . . .*, (June):28–36.

[Barabasi and Oltvai, 2004] Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–113.

[Barbehenn, 1998] Barbehenn, M. (1998). A note on the complexity of dijkstra's algorithm for graphs with weighted vertices. *IEEE Trans. Comput.*, 47(2):263–.

[Barnickel et al., 2009] Barnickel, T., Weston, J., Collobert, R., Mewes, H.-W. W., and Stümpflen, V. (2009). Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS One*, 4(7):e6393.

[Baud et al., 2002] Baud, S., Boutin, J.-P., Miquel, M., Lepiniec, L., and Rochat, C. (2002). An integrated overview of seed development in arabidopsis thaliana ecotype WS. *Plant Physiol. Biochem.*, 40(2):151–160.

[Baud et al., 2009] Baud, S., Wuillème, S., To, A., Rochat, C., and Lepiniec, L. (2009). Role of WRINKLED1 in the transcriptional regulation of glycolytic and fatty acid biosynthetic genes in arabidopsis. *Plant J.*, 60(6):933–947.

[Bellinger et al., 2004] Bellinger, G., Castro, D., and Mills, A. (2004). Data, information, knowledge, and wisdom.

[Bengio et al., 2001] Bengio, Y., Ducharme, R., and Vincent, P. (2001). A neural probabilistic language model. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 932–938. MIT Press.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

[Bentsink et al., 2008] Bentsink, L., Leónie, B., and Maarten, K. (2008). Seed dormancy and germination. *Arabidopsis Book*, 6:e0119.

[Bewley, 1997] Bewley, J. D. (1997). Seed germination and dormancy. *The plant cell*, 9(7):1055.

[Bird and Liberman, 2001] Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Commun.*, 33(1):23–60.

[Blais and Dynlacht, 2005] Blais, A. and Dynlacht, B. D. (2005). Constructing transcriptional regulatory networks. *Genes Dev.*, 19(13):1499–1511.

[Blaschke et al., 1999] Blaschke, C., Andrade, M. A., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, pages 60–67.

[Blaschke et al., 2001] Blaschke, C., Oliveros, J. C., and Valencia, A. (2001). Mining functional information associated with expression arrays. *Funct. Integr. Genomics*, 1(4):256–268.

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–70.

[Bonneau-Maynard et al., 2005] Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., and Mostefa, D. (2005). Semantic annotation of the french media dialog corpus. In *Ninth European Conference on Speech Communication and Technology.*

[Bossy et al., 2013a] Bossy, R., Bessières, P., and Nédellec, C. (2013a). Bionlp shared task 2013–an overview of the genic regulation network task. *Proceedings of the BioNLP 2013 Workshop*, page 153.

[Bossy et al., 2013b] Bossy, R., Golik, W., Ratkovic, Z., Bessières, P., and Nédellec, C. (2013b). BioNLP shared task 2013–an overview of the bacteria biotope task. *Proceedings of the BioNLP Workshop at ACL Conference.*

[Bossy et al., 2015] Bossy, R., Golik, W., Ratkovic, Z., Valsamou, D., Bessières, P., and Nédellec, C. (2015). Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task. *BMC Bioinformatics*, 16(Suppl 10):S1.

[Bossy et al., 2011a] Bossy, R., Jourde, J., Bessières, P., Van De Guchte, M., and Nédellec, C. (2011a). Bionlp shared task 2011: bacteria biotope. In *Proceedings of the BioNLP Workshop at ACL Conference*, pages 56–64.

[Bossy et al., 2011b] Bossy, R., Jourde, J., Bessières, P., van de Guchte, M., and Nédellec, C. (2011b). BioNLP shared task 2011: bacteria biotope. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 56–64. Association for Computational Linguistics.

[Bossy et al., 2012] Bossy, R., Jourde, J., Manine, A.-P., Veber, P., Alphonse, E., Van De Guchte, M., Bessières, P., and Nédellec, C. (2012). BioNLP shared Task-The bacteria track. *BMC Bioinformatics*, 13(Suppl 11):S3.

[Boughorbel et al., 2004] Boughorbel, S., Tarel, J.-P., and Fleuret, F. (2004). Non-Mercer kernels for SVM object recognition. *Bmvc.*

[Brady et al., 2003] Brady, S. M., Sarkar, S. F., Bonetta, D., and McCourt, P. (2003). The ABSCISIC ACID INSENSITIVE 3 (ABI3) gene is modulated by farnesylation and is involved in auxin signaling and lateral root development in arabidopsis. *Plant J.*, 34(1):67–75.

[Braybrook et al., 2006] Braybrook, S. A., Stone, S. L., Park, S., Bui, A. Q., Le, B. H., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (2006). Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis. *Proc. Natl. Acad. Sci. U. S. A.*, 103(9):3468–3473.

[Bretonnel Cohen and Hunter, 2004] Bretonnel Cohen, K. and Hunter, L. (2004). Natural language processing and systems biology. In *Artificial Intelligence Methods And Tools For Systems Biology*, Computational Biology, pages 147–173. Springer Netherlands.

[Brin et al., 1998] Brin, S., Motwani, R., Page, L., and Winograd, T. (1998). What can you do with a web in your pocket? *IEEE Data Eng. Bull.*, 21(2):37–47.

[Brocard-Gifford et al., 2003] Brocard-Gifford, I. M., Lynch, T. J., and Finkelstein, R. R. (2003). Regulatory networks in seeds integrating developmental, abscisic acid, sugar, and light signaling. *Plant Physiol.*, 131(1):78–92.

[Buitelaar, 2009] Buitelaar, P. (2009). Natural language processing for the semantic web.

[Buitelaar and Cimiano, 2008] Buitelaar, P. and Cimiano, P. (2008). *Ontology learning and population: bridging the gap between text and knowledge*, volume 167.

[Buitelaar and Magnini, 2005] Buitelaar, P. and Magnini, B. (2005). Ontology learning from text: An overview. In Buitelaar, P., Cimiano, P., and Magnini, B., editors, *Ontology Learning from Text: Methods, Applications and Evaluation*, pages 3–12. IOS Press.

[Buitelaar et al., 2004] Buitelaar, P., Paul, B., Daniel, O., and Michael, S. (2004). A protégé Plug-In for ontology extraction from text based on linguistic analysis. In *Lecture Notes in Computer Science*, pages 31–44.

[Bunescu and Mooney, 2005] Bunescu and Mooney (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of Empricial Methods in Natural Language Processing*, EMNLP '05, pages 724–731.

[Casson and Lindsey, 2006] Casson, S. A. and Lindsey, K. (2006). The turnip mutant of arabidopsis reveals that LEAFY COTYLEDON1 expression mediates the effects of auxin and sugars to promote embryonic cell identity. *Plant Physiol.*, 142(2):526–541.

[Cernac et al., 2006] Cernac, A., Andre, C., Hoffmann-Benning, S., and Benning, C. (2006). WRI1 is required for seed germination and seedling establishment. *Plant Physiol.*, 141(2):745–757.

[Chaix, 2015] Chaix, E. (2015). Information extraction challenge gene regulation network in arabidopsis thaliana (GRNA).

[Chaix et al., 2016a] Chaix, E., Dubreucq, B., Fatihi, A., Valsamou, D., Bossy, R., Ba, M., Deléger, L., Zweigenbaum&, P., Bessieres, P., Lepiniec, L., et al. (2016a). Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop - Association for Computational Linguistics*, pages 1–11.

[Chaix et al., 2015a] Chaix, E., Dubreucq, B., Valsamou, D., Fatihi, A., Bossy, R., Deléger, L., Zweigenbaum, P., Bessières, P., Lepiniec, L., and Nédellec, C. (2015a). A knowledge model of regulatory networks involved in arabidopsis seed development for information extraction and integration from text. In *BioCreative 5 Challenge Workshop, At Sevilla, Spain.*

[Chaix et al., 2015b] Chaix, E., Dubreucq, B., Valsamou, D., Fatihi, A., Deléger, L., Bossy, R., Zweigenbaum, P., Bessières, P., Lepiniec, L., and Nédellec, C. (2015b). Instance of annotation guidelines v 4.6. https://docs.google.com/document/d/1Zj3s5T30Lhfu4AHNdrCNaVSR0ZLxyeMQI_oHIlZt9zM/edit?usp=sharing.

[Chaix et al., 2016b] Chaix, E., Dubreucq, B., Valsamou, D., Fatihi, A., Deléger, L., Bossy, R., Zweigenbaum, P., Bessières, P., Lepiniec, L., and Nédellec, C. (2016b). *Annotation Guidelines BIONLP-ST 2016 - SeeDev task.* INRA.

[Chaix et al., 2016c] Chaix, E., Dubreucq, B., Valsamou, D., Fatihi, A., Deléger, L., Bossy, R., Zweigenbaum, P., Bessières, P., Lepiniec, L., and Nédellec, C. (2016c). BIONLP-ST 2016 - SeeDev task - signature of relation arguments. https://docs.google.com/spreadsheets/d/19cnRx0WFFPBVAnPMaxjCmEYSaaTjWSyMDgxajOTFAVs/edit.

[Chamberlain et al., 2008] Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pages 42–49.

[Che et al., 2005] Che, W., Jiang, J., Su, Z., Pan, Y., and Liu, T. (2005). Improved-edit-distance kernel for chinese relation extraction. *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'05)*, pages 134–139.

[Chen and Manning, 2014] Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. *EMNLP.*

[Chinchor et al., 1993] Chinchor, N., Lewis, D. D., and Hirschman, L. (1993). Evaluating message understanding systems: An analysis of the third message understanding conference (MUC-3). *Comput. Linguist.*, 19(3):409–449.

[Chowdhury and Lavelli, 2011] Chowdhury, M. F. M. and Lavelli, A. (2011). Drug-drug interaction extraction using composite kernels. *Challenge Task on Drug-Drug Interaction Extraction*, pages 27–33.

[Chun et al., 2006] Chun, H. W., Tsuruoka, Y., and Kim, J.-D. (2006). Extraction of gene-disease relations from medline using domain dictionaries and machine learning. *Pac Symp . . .*, 15:4–15.

[Clark and Curran, 2007] Clark, S. and Curran, J. R. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Comput. Linguist.*, 33(4):493–552.

[Claveau, 2013] Claveau, V. (2013). IRISA participation to BioNLP-ST 2013 : lazy-learning and information retrieval for information extraction tasks. pages 188–196.

[Cohen, 2010] Cohen, K. B. (2010). BioNLP: Biomedical text mining. In *Handbook of Natural Language Processing*, Machine Learning & Pattern Recognition, pages 605–625. Chapman & Hall, Boca Raton, 2nd edition.

[Cohen et al., 2005] Cohen, K. B., Ogren, P. V., Fox, L., and Hunter, L. (2005). Empirical data on corpus design and usage in biomedical natural language processing. *AMIA Annu. Symp. Proc.*, pages 156–160.

[Cohen et al., 2003] Cohen, W., Ravikumar, P., and Fienberg, S. (2003). A comparison of string metrics for matching names and records. cs.cmu.edu.

[Collier and Takeuchi, 2004] Collier, N. and Takeuchi, K. (2004). Comparison of character-level and part of speech features for name recognition in biomedical texts. *J. Biomed. Inform.*, 37(6):423–435.

[Craven and Kumlien, 1999] Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, pages 77–86.

[Culotta et al., 2006] Culotta, A., Mccallum, A., Betz, J., Bush, N., and Prescott, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics.*

[Culotta and Sorensen, 2004] Culotta, A. and Sorensen, J. (2004). Dependency tree kernels for relation extraction. *Proceedings of the 42nd Annual Meeting on . . .*, 4(Table 2):423–es.

[Cunningham, 2002] Cunningham, H. (2002). GATE, a general architecture for text engineering. *Comput. Hum.*, 36(2):223–254.

[Cunningham, 2006] Cunningham, H. (2006). Information extraction, automatic. In *Encyclopedia of Language & Linguistics*, volume 5, pages 665–677.

[Curaba et al., 2004] Curaba, J., Moritz, T., Blervaque, R., Parcy, F., Raz, V., Herzog, M., and Vachon, G. (2004). AtGA3ox2, a key gene responsible for bioactive gibberellin biosynthesis, is regulated during embryogenesis by LEAFY COTYLEDON2 and FUSCA3 in arabidopsis. *Plant Physiol.*, 136(3):3660–3669.

[Dandapat et al., 2009] Dandapat, S., Sandipan, D., Priyanka, B., Monojit, C., and Kalika, B. (2009). Complex linguistic annotation — no easy way out! In *Proceedings of the Third Linguistic Annotation Workshop on - ACL-IJCNLP '09.*

[Daraselia et al., 2004] Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–611.

[Davuluri et al., 2003] Davuluri, R. V., Sun, H., Palaniswamy, S. K., Matthews, N., Molina, C., Kurtz, M., and Grotewold, E. (2003). Agris: Arabidopsis gene regulatory information server, an information resource of arabidopsis cis-regulatory elements and transcription factors. *BMC bioinformatics*, 4(1):25.

[De Marneffe et al., 2006] De Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Language Resources and Evaluation Conference*, volume 6 of *LREC '06*, pages 449–454.

[Dice, 1945] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

[Doddington et al., 2004] Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ACE) Program-Tasks, data, and evaluation. In *LREC*. ldc.upenn.edu.

[Dérozier et al., 2011] Dérozier, S., Samson, F., Tamby, J.-P., Guichard, C., Brunaud, V., Grevet, P., Gagnot, S., Label, P., Leplé, J.-C., Lecharny, A., et al. (2011). Exploration of plant genomes in the flagdb++ environment. *Plant methods*, 7(1):1.

[Dubreucq, 2014] Dubreucq, B. (2014). Elaboration du réseau de régulation impliqué dans le développement de la graine chez arabidopsis thaliana.

[Dubreucq et al., 2009] Dubreucq, B., Baud, S., Debeaujon, I., Dubos, C., Marion-Poll, A., Miquel, M., North, H., Rochat, C., J.-M., R., and Lepiniec, L. (2009). Seed development. In *Plant Developmental Biology - Biotechnological Perspectives*, pages 341–359. Springer.

[Dubreucq et al., 2015] Dubreucq, B., Valsamou, D., Fatihi, A., Chaix, E., Bossy, R., Bessieres, P., Deléger, L., Zweigenbaum, P., Nédellec, C., and Lepiniec, L. (2015). Information extraction from articles for the elaboration of the regulatory networks involved in arabidopsis seed development. In *The 26th International conference on Arabidopsis research, At Paris, France.*

[Eilbeck et al., 1999] Eilbeck, K., Brass, A., Paton, N., and Hodgman, C. (1999). INTERACT: an object oriented protein-protein interaction database. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, pages 87–94.

[Erkan et al., 2007] Erkan, G., Ozgur, A., and Radev, D. R. (2007). Extracting interacting protein pairs and evidence sentences by using dependency parsing and machine learning techniques. *Proceedings of the Second BioCreative Challenge Workshop*, 1:228–237.

[Espinosa-Soto et al., 2004] Espinosa-Soto, C., Padilla-Longoria, P., and Alvarez-Buylla, E. R. (2004). A gene regulatory network model for Cell-Fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *Plant Cell*, 16(11):2923–2939.

[Fang and Fernandez, 2002] Fang, S.-C. and Fernandez, D. E. (2002). Effect of regulated overexpression of the MADS domain factor AGL15 on flower senescence and fruit maturation. *Plant Physiol.*, 130(1):78–89.

[Faro et al., 2012] Faro, A., Giordano, D., and Spampinato, C. (2012). Combining literature text mining with microarray data: Advances for system biology modeling. *Brief. Bioinform.*, 13(1):61–82.

[Fayruzov et al., 2009] Fayruzov, T., De Cock, M., Cornelis, C., and Hoste, V. (2009). Linguistic feature analysis for protein interaction extraction. *BMC Bioinformatics*, 10:374.

[Federhen, 2012] Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Res.*, 40(Database issue):D136–43.

[Fellbaum, 1998] Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

[Fernandez et al., 2000] Fernandez, D. E., Heck, G. R., Perry, S. E., Patterson, S. E., Bleecker, A. B., and Fang, S. C. (2000). The embryo MADS domain factor AGL15 acts postembryonically. inhibition of perianth senescence and abscission via constitutive expression. *Plant Cell*, 12(2):183–198.

[Finkel et al., 2005a] Finkel, J. R., Grenager, T., and Manning, C. (2005a). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

[Finkel et al., 2005b] Finkel, J. R., Grenager, T., and Manning, C. (2005b). Incorporating non-local information into information extraction systems by gibbs sampling. *Acl*, (1995):363–370.

[Fort, 2011] Fort, K. (2011). Corpus linguistics (lecture notes).

[Fort et al., 2011] Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Comput. Linguist.*, 37(2):413–420.

[Freitag and McCallum, 1999] Freitag, D. and McCallum, A. (1999). Information extraction with HMMs and shrinkage. In *Proceedings of the AAAI-99 workshop on machine learning for information extraction*, pages 31–36.

[Friedman et al., 2001] Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(suppl 1):S74–S82.

[Fukuda et al., 1998] Fukuda, K.-I., Tsunoda, T., Tamura, A., and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. In *Pacific Symposium on Biocomputing*, pages 707–718.

[Fundel et al., 2007] Fundel, K., Küffner, R., and Zimmer, R. (2007). RelEx - relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

[Gabbard et al., 2011] Gabbard, R., Freedman, M., and Weischedel, R. (2011). Coreference for learning to extract relations: yes, virginia, coreference matters. *... of the 49th Annual Meeting of ...*, pages 288–293.

[Galibert et al., 2010] Galibert, O., Quintard, L., Rosset, S., Zweigenbaum, P., Nédellec, C., Aubin, S., Gillard, L., Raysz, J.-P., Pois, D., and Others (2010). Named and specific entity detection in varied data: The quæro named entity baseline evaluation.

[Gardner et al., 2013] Gardner, M., Talukdar, P. P., Kisiel, B., and Mitchell, T. (2013). Improving learning and inference in a large knowledge-base using latent syntactic cues.

[Garside et al., 1997] Garside, R., Leech, G. N., and McEnery, T. (1997). *Corpus annotation: linguistic information from computer text corpora*. Taylor & Francis.

[Gazzarrini et al., 2004] Gazzarrini, S., Tsuchiya, Y., Lumba, S., Okamoto, M., and McCourt, P. (2004). The transcription factor FUSCA3 controls developmental timing in arabidopsis through the hormones gibberellin and abscisic acid. *Dev. Cell*, 7(3):373–385.

[Gieger et al., 2003] Gieger, C., Deneke, H., and Fluck, J. (2003). The future of text mining in genome-based clinical research. *BIOSILICO*, 1(3):97–102.

[Giuliano et al., 2006] Giuliano, C., Lavelli, A., and Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the European chapter of the Association for Computational Linguistics Conference*, volume 2006 of *EACL '06*, pages 98–113.

[Goelzer et al., 2008] Goelzer, A., Bekkal Brikci, F., Martin-Verstraete, I., Noirot, P., Bessières, P., Aymerich, S., and Fromion, V. (2008). Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of bacillus subtilis. *BMC Syst. Biol.*, 2(1):20.

[Goldberg and Levy, 2014] Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method.

[Golik et al., 2013] Golik, W., Bossy, R., Ratkovic, Z., and Claire, N. (2013). Improving term extraction with linguistic analysis in the biomedical domain. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing13), Special Issue of the journal Research in Computing Science*, pages 24–30. cicling.org.

[Golik et al., 2012a] Golik, W., Dameron, O., Bugeon, J., Fatet, A., Hue, I., Hurtaud, C., Reichstadt, M., Salaün, M.-C., Vernet, J., and Joret, L. (2012a). ATOL: the multi-species livestock trait ontology. In *Metadata and Semantics Research*, pages 289–300. Springer.

[Golik et al., 2012b] Golik, W., Dameron, O., Bugeon, J., Fatet, A., Hue, I., and others (2012b). ATOL: the multi-species livestock trait ontology. *Metadata and Semantics.*

[Golik et al., 2011] Golik, W., Warnier, P., and Nédellec, C. (2011). Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence*, pages 37–39.

[Google, 2013] Google (2013). Relation extraction corpus. https://code.google.com/p/relation-extraction-corpus/.

[Gormley et al., 2014] Gormley, M. R., Dredze, M., Yu, M., Intelligence, M., and Processing, S. (2014). Combining word embeddings and feature embeddings for fine-grained relation extraction.

[Grishman, 1997] Grishman, R. (1997). Information extraction: Techniques and challenges. In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, Lecture Notes in Computer Science, pages 10–27. Springer Berlin Heidelberg.

[Grishman, 2003] Grishman, R. (2003). Information extraction. In Clark, A., Fox, C., and and Shalom Lappin, editors, *The Handbook of Computational Linguistics and Natural Language Processing*, chapter 18, pages 517–530. Wiley-Blackwell.

[Grishman, 2012] Grishman, R. (2012). Information extraction: Capabilities and challenges. *Notes prepared for the 2012 International Winter School in Language and Speech Technologies.*

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. *Proceedings of the 16th conference on Computational linguistics*, 1:466–471.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

[Guarino and Welty, 2000] Guarino, N. and Welty, C. (2000). A formal ontology of properties. In *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, Lecture Notes in Computer Science, pages 97–112. Springer Berlin Heidelberg.

[GuoDong et al., 2005] GuoDong, Z., Jian, S., Jie, Z., and Min, Z. (2005). Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 427–434. Association for Computational Linguistics.

[Gut and Bayerl, 2004] Gut, U. and Bayerl, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Speech Prosody 2004, International Conference.*

[Haasdonk, 2005] Haasdonk, B. (2005). Feature space interpretation of SVMs with indefinite kernels. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(4):482–492.

[Hakenberg et al., 2004] Hakenberg, J., Schmeier, S., Kowald, A., Klipp, E., and Leser, U. (2004). Finding kinetic parameters using text mining. *OMICS*, 8(2):131–152.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

[Harris, 1954] Harris, Z. S. (1954). Distributional structure. *Word World.*

[Hastie et al., 2001a] Hastie, T., Friedman, J., and Tibshirani, R. (2001a). Support vector machines and flexible discriminants. In *The Elements of Statistical Learning*, Springer Series in Statistics, pages 371–409. Springer New York.

[Hastie et al., 2001b] Hastie, T., Tibshirani, R., and Friedman, J. (2001b). *The elements of statistical learning*, volume 1 of *Springer series in statistics*. Springer, Berlin, second edition.

[He et al., 2013] He, L., Yang, Z., Zhao, Z., Lin, H., and Li, Y. (2013). Extracting drug-drug interaction from the biomedical literature using a stacked generalization-based approach. *PLoS One*, 8(6):e65814.

[Hearst, 1999] Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics -*.

[Hersh and William, 2004] Hersh, W. and William, H. (2004). Report on TREC 2003 genomics track first-year results and future plans. *ACM SIGIR Forum*, 38(1):69.

[Hersh et al., 2008] Hersh, W., William, H., and Ellen, V. (2008). TREC genomics special issue overview. *Inf. Retr. Boston.*, 12(1):1–15.

[Hersh et al., 2006] Hersh, W. R., Cohen, A. M., Roberts, P. M., and Rekapalli, H. K. (2006). TREC 2006 genomics track overview. In *TREC*. skynet.ohsu.edu.

[Hill et al., 2008] Hill, K., Wang, H., and Perry, S. E. (2008). A transcriptional repression motif in the MADS factor AGL15 is involved in recruitment of histone deacetylase complex components. *Plant J.*, 53(1):172–185.

[Hirschman et al., 2002a] Hirschman, L., Morgan, A. A., and Yeh, A. S. (2002a). Rutabaga by any other name: extracting biological names. *J. Biomed. Inform.*, 35(4):247–259.

[Hirschman et al., 2002b] Hirschman, L., Park, J. C., Tsujii, J., Wong, L., and Wu, C. H. (2002b). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561.

[Hirschman et al., 2005] Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1.

[Hirst and St-Onge, 1998] Hirst, G. and St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332.

[Hoffmann and Valencia, 2004] Hoffmann, R. and Valencia, A. (2004). A gene network for navigating the literature. *Nat. Genet.*, 36(7):664.

[Hsuan-tien Lin, 2003] Hsuan-tien Lin, C.-J. L. (2003). A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods.

[Huffman, 1995] Huffman, S. (1995). Learning information extraction patterns from examples. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing.*

[Ideker et al., 2003] Ideker, T., Galitski, T., and Hood, L. (2003). A NEW APPROACH TO DECODING LIFE: Systems biology.

[Imoto et al., 2011] Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2011). COMBINING MICROARRAYS AND BIOLOGICAL KNOWLEDGE FOR ESTIMATING GENE NETWORKS VIA BAYESIAN NETWORKS. *J. Bioinform. Comput. Biol.*

[Jaro, 1989] Jaro, M. A. (1989). Advances in Record-Linkage methodology as applied to matching the 1985 census of tampa, florida. *J. Am. Stat. Assoc.*, 84(406):414–420.

[Jaro, 1995] Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Stat. Med.*, 14(5-7):491–498.

[Jensen et al., 2006] Jensen, L. J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, 7(2):119–129.

[Jenssen et al., 2001] Jenssen, T.-K., Lægreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, 28(1):21–28.

[Jiang and Zhai, 2007] Jiang, J. and Zhai, C. (2007). An empirical study of tokenization strategies for biomedical information retrieval. *Inf. Retr. Boston.*, 10(4-5):341–363.

[Jiang and Conrath, 1997] Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy.

[Jin et al., 2013] Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2013). Planttfdb 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic acids research*, page gkt1016.

[Kagaya et al., 2005a] Kagaya, Y., Okuda, R., Ban, A., Toyoshima, R., Tsutsumida, K., Usui, H., Yamamoto, A., and Hattori, T. (2005a). Indirect ABA-dependent regulation of seed storage protein genes by FUSCA3 transcription factor in arabidopsis. *Plant Cell Physiol.*, 46(2):300–311.

[Kagaya et al., 2005b] Kagaya, Y., Toyoshima, R., Okuda, R., Usui, H., Yamamoto, A., and Hattori, T. (2005b). LEAFY COTYLEDON1 controls seed storage protein genes through its regulation of FUSCA3 and ABSCISIC ACID INSENSITIVE3. *Plant Cell Physiol.*, 46(3):399–406.

[Kambhatla, 2004] Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations.

[Karlova et al., 2006] Karlova, R., Boeren, S., Russinova, E., Aker, J., Vervoort, J., and de Vries, S. (2006). The arabidopsis SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE1 protein complex includes BRASSINOSTEROID-INSENSITIVE1. *Plant Cell*, 18(3):626–638.

[Kauffman, 1969] Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, 22(3):437–467.

[Kayed et al., 2006] Kayed, M., Girgis, M. R. M. R., Shaalan, K. F. K. F., Chang, C.-H., Kayed, M., Girgis, M. R. M. R., and Shaalan, K. F. K. F. (2006). A survey of web information extraction systems. *IEEE Trans. Knowl. Data Eng.*, 18(10):1411–1428.

[Kazama et al., 2002] Kazama, J., Jun'ichi, K., Takaki, M., Yoshihiro, O., and Jun'ichi, T. (2002). Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain -*.

[Kim et al., 2011] Kim, Wang, Y., Takagi, and Yonezawa (2011). Overview of genia event task in BioNLP shared task 2011. *ACL HLT 2011*, page 7.

[Kim et al., 2004] Kim, J.-D., Jin-Dong, K., Tomoko, O., Yoshimasa, T., Yuka, T., and Nigel, C. (2004). Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA '04*.

[Kim et al., 2012] Kim, J.-D., Nguyen, N., Wang, Y., Tsujii, J., Takagi, T., and Yonezawa, A. (2012). The genia event and protein coreference tasks of the BioNLP shared task 2011. *BMC Bioinformatics*, 13 Suppl 1(Suppl 11):S1.

[Kim et al., 2009a] Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J. (2009a). Overview of BioNLP'09 shared task on event extraction. In *Proceedings of BioNLP Workshop at ACL Conference*.

[Kim et al., 2009b] Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J. (2009b). Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Kim et al., 2003] Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

[Kim et al., 2008] Kim, J.-D., Ohta, T., and Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10.

[Kim et al., 2011] Kim, J.-D., Pyysalo, S., Ohta, T., and Bossy, R. (2011). Overview of bionlp shared task 2011. *... BioNLP Shared Task ...*, 2009:1–6.

[Kim et al., 2015] Kim, J.-D., Wang, Y., and Nakajima, S. (2015). TextAE. http://textae.pubannotation.org/docs/about/.

[Kim et al., 2013a] Kim, J. D., Wang, Y., and Yasunori, Y. (2013a). The genia event extraction shared task, 2013 edition-overview. *the BioNLP Shared Task 2013 ...*.

[Kim et al., 2013b] Kim, S., Iglesias-Sucasas, M., and Viollier, V. (2013b). The FAO geopolitical ontology: A reference for Country-Based information. *Journal of Agricultural & Food Information*, 14(1):50–65.

[Kolb, 2008] Kolb, P. (2008). Disco: A multilingual database of distributionally similar words. *Proceedings of KONVENS-2008, Berlin*, (2003).

[Kolb, 2009] Kolb, P. (2009). Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic Conference on Computational Linguistics-NODALIDA'09*.

[Krallinger et al., 2005] Krallinger, M., Erhardt, R., and Valencia, A. (2005). Text-mining approaches in molecular biology and biomedicine. *Drug Discov. Today*, 10(6):439–445.

[Krallinger et al., 2009] Krallinger, M., Rojas, A. M., and Valencia, A. (2009). Creating reference datasets for systems biology applications using text mining. *Ann. N. Y. Acad. Sci.*, 1158:14–28.

[Krouk et al., 2013] Krouk, G., Lingeman, J., Colon, A. M., Coruzzi, G., and Shasha, D. (2013). Gene regulatory networks in plants: learning causality from time and perturbation. *Genome Biol.*, 14(6):123.

[Kulick et al., 2004] Kulick, S., Bies, A., and Liberman, M. (2004). Integrated annotation for biomedical information extraction. *Proc. of the Human ...*, pages 61–68.

[Kwong et al., 2003] Kwong, R. W., Bui, A. Q., Lee, H., Kwong, L. W., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (2003). LEAFY COTYLEDON1-LIKE defines a class of regulators essential for embryo development. *Plant Cell*, 15(1):5–18.

[Lang and Davis, 2010] Lang, J. and Davis, E. (2010). Redmine-open source project management web-application.

[Lara et al., 2003] Lara, P., Oñate Sánchez, L., Abraham, Z., Ferrándiz, C., Díaz, I., Carbonero, P., and Vicente-Carbajosa, J. (2003). Synergistic activation of seed storage protein gene expression in arabidopsis by ABI3 and two bZIPs related to OPAQUE2. *J. Biol. Chem.*, 278(23):21003–21011.

[LDC, 1992] LDC (1992). Linguistic data consortium - linguistic data consortium. https://catalog.ldc.upenn.edu/.

[LDC, 1993] LDC (1993). TIPSTER complete - linguistic data consortium. https://catalog.ldc.upenn.edu/LDC93T3A.

[LDC, 2001] LDC (2001). Message understanding conference (MUC) 7 - linguistic data consortium. https://catalog.ldc.upenn.edu/LDC2001T02.

[LDC, 2008a] LDC (2008a). The new york times annotated corpus - linguistic data consortium. https://catalog.ldc.upenn.edu/LDC2008T19.

[LDC, 2008b] LDC (2008b). PennBioIE oncology 1.0 - linguistic data consortium. https://catalog.ldc.upenn.edu/LDC2008T21.

[LDC, 2011a] LDC (2011a). OntoNotes release 4.0 - linguistic data consortium. https://catalog.ldc.upenn.edu/LDC2011T03.

[LDC, 2011b] LDC (2011b). SemEval-2010 task 1 OntoNotes english: Coreference resolution in multiple languages - linguistic data consortium. https://catalog.ldc.upenn.edu/LDC2011T01.

[Leacock and Chodorow, 1998] Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

[Lee et al., 2003] Lee, H., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (2003). Arabidopsis LEAFY COTYLEDON1 represents a functionally specialized subunit of the CCAAT binding transcription factor. *Proc. Natl. Acad. Sci. U. S. A.*, 100(4):2152–2156.

[Lee et al., 2008] Lee, J. M., Min Lee, J., Gianchandani, E. P., Eddy, J. A., and Papin, J. A. (2008). Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput. Biol.*, 4(5):e1000086.

[Leech, 1993] Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing*, 8(4):275–281.

[Lepiniec et al., 2006] Lepiniec, L., Debeaujon, I., Routaboul, J.-M., Baudry, A., Pourcel, L., Nesi, N., and Caboche, M. (2006). Genetics and biochemistry of seed flavonoids. *Annu. Rev. Plant Biol.*, 57:405–430.

[Levy and Goldberg, 2014] Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

[Lewis, 1991] Lewis, D. D. (1991). Data extraction as text categorization: an experiment with the MUC-3 corpus. In *Proceedings of the 3rd conference on Message understanding*, pages 245–255. Association for Computational Linguistics.

[Lewis et al., 2004] Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

[Li et al., 2013] Li, C., Liakata, M., and Rebholz-Schuhmann, D. (2013). Biological network extraction from scientific literature: state of the art and challenges. *Brief. Bioinform.*

[Li et al., 2011] Li, Q., Anzaroot, S., Lin, W. P., and Li, X. (2011). Joint inference for cross-document information extraction. *conference on Information and.*

[Liao and Noble, 2002] Liao, L. and Noble, W. S. (2002). Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In *Proceedings of the Sixth Annual International Conference on Computational Biology*, RECOMB '02, pages 225–232, New York, NY, USA. ACM.

[Liao and Noble, 2003] Liao, L. and Noble, W. S. (2003). Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.*, 10(6):857–868.

[Liao et al., 2010] Liao, S., Grishman, R., York, N., and Grishman, R. (2010). Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '08, pages 789–797.

[Lin, 1998] Lin, D. (1998). An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304.

[Lin et al., 2004] Lin, Y.-H., Liang, T., and Hsinehu, T. (2004). Pronominal and sortal anaphora resolution for biomedical literature. In *ROCLING*.

[Lingren et al., 2014] Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., and Solti, I. (2014). Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *J. Am. Med. Inform. Assoc.*, 21(3):406–413.

[Linguistic Data Consortium, 1994] Linguistic Data Consortium (1994). ECI multilingual text - linguistic data consortium. https://catalog.ldc.upenn.edu/LDC94T5.

[Liu et al., 2012] Liu, B., Qian, L., Zhou, G., and Zhu, Q. (2012). Exploiting dependency information for Feature-Based Protein-Protein interaction extraction. *of the 2011, International Conference on.*

[Liu and Tamer Özsu, 2009] Liu, L. and Tamer Özsu, M. (2009). *Encyclopedia of Database Systems*. Springer.

[Liu et al., 2007] Liu, Y., Shi, Z., and Sarkar, A. (2007). Exploiting rich syntactic information for relation extraction from biomedical articles. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers on XX*, pages 97–100. Association for Computational Linguistics.

[Lotan et al., 1998] Lotan, T., Ohto, M., Yee, K. M., West, M. A., Lo, R., Kwong, R. W., Yamagishi, K., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (1998). Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell*, 93(7):1195–1205.

[Lüdeling, 2008] Lüdeling, A. (2008). *Corpus linguistics*, volume 1. Walter de Gruyter.

[Luerssen et al., 1998] Luerssen, H., Kirik, V., Herrmann, P., and Miséra, S. (1998). FUSCA3 encodes a protein with a conserved VP1/AB13-like B3 domain which is of functional importance for the regulation of seed maturation in arabidopsis thaliana. *Plant J.*, 15(6):755–764.

[Maedche and Staab, 2001] Maedche, A. and Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intell. Syst.*, (2):72–79.

[Maeo et al., 2009] Maeo, K., Tokuda, T., Ayame, A., Mitsui, N., Kawai, T., Tsukagoshi, H., Ishiguro, S., and Nakamura, K. (2009). An AP2-type transcription factor, WRINKLED1, of arabidopsis thaliana binds to the AW-box sequence conserved among proximal upstream regions of genes involved in fatty acid synthesis. *Plant J.*, 60(3):476–487.

[Maes et al., 2001] Maes, T., Van de Steene, N., Zethof, J., Karimi, M., D'Hauw, M., Mares, G., Van Montagu, M., and Gerats, T. (2001). Petunia ap2-like genes and their role in flower and seed development. *Plant Cell*, 13(2):229–244.

[Mahoney, 2011] Mahoney, M. (2011). enwiki9- about the test data. http://mattmahoney.net/dc/textdata.html.

[Manine et al., 2008] Manine, A. P., Alphonse, E., and Bessières, P. (2008). Information extraction as an ontology population task and its application to genic interactions. In *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, volume 2, pages 74–81.

[Manine et al., 2009] Manine, A.-P., Alphonse, E., and Bessières, P. (2009). Learning ontological rules to extract multiple relations of genic interactions from text. *Int. J. Med. Inform.*, 78(12):e31–8.

[Marcus et al., 1993] Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330.

[Martin, 2008] Martin, R. C. (2008). *Clean Code: A Handbook of Agile Software Craftsmanship.*

[McCallum, 2002] McCallum, A. (2002). Efficiently inducing features of conditional random fields. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 403–410. Morgan Kaufmann Publishers Inc.

[McCallum, 2005] McCallum, A. (2005). Information extraction. *Queueing Syst.*, 3(9):48.

[McDonald et al., 2004] McDonald, D. M., Chen, H., Su, H., and Marshall, B. B. (2004). Extracting gene pathway relations using a hybrid grammar: the arizona relation parser. *Bioinformatics*, 20(18):3370–3378.

[McDonald and Pereira, 2005] McDonald, R. and Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, 6 Suppl 1:S6.

[Mendoza et al., 1999] Mendoza, L., Thieffry, D., and Alvarez-Buylla, E. R. (1999). Genetic control of flower morphogenesis in arabidopsis thaliana: a logical analysis. *Bioinformatics*, 15(7):593–606.

[Mikolov et al., 2013a] Mikolov, T., Sutskever, I., and Chen, K. (2013a). Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.*, pages 1–9.

[Mikolov et al., 2013b] Mikolov, T., View, M., Corrado, G., Chen, K., and Dean, J. (2013b). Efficient estimation of word representations in vector space. pages 1–12.

[Mikolov et al., 2013c] Mikolov, T., Yih, W.-T., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. *Proceedings of NAACL-HLT.*

[Miller et al., 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database*. *International Journal of Lexicography*, 3(4):235–244.

[Mingrui Wu et al., 2006] Mingrui Wu, Bernhard Scholkopf, and Gokhan Bakir (2006). A direct method for building sparse kernel learning algorithms. *J. Mach. Learn. Res.*, 7:603–624.

[Minh et al., 2006] Minh, H. Q., Niyogi, P., and Yao, Y. (2006). Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pages 154–168.

[Mintz et al., 2009] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural*

*Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

[Miwa et al., 2009] Miwa, M., Sætre, R., Miyao, Y., and Tsujii, J. (2009). Protein–protein interaction extraction by leveraging multiple kernels and parsers. *Int. J. Med. Inform.*, 78(12):e39–e46.

[Miwa and Sasaki, 2014] Miwa, M. and Sasaki, Y. (2014). Modeling joint entity and relation extraction with table representation. In *EMNLP 2014*, pages 1858–1869.

[Miyao et al., 2003] Miyao, Y., Ninomiya, T., and Tsujii, J. (2003). Probabilistic modeling of argument structures including non-local dependencies. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, RANLP '03, pages 285–291.

[Miyao et al., 2009] Miyao, Y., Sagae, K., Saetre, R., Matsuzaki, T., and Tsujii, J. (2009). Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics*, 25(3):394–400.

[Miyao and Tsujii, 2005] Miyao, Y. and Tsujii, J. (2005). Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics Conference*, ACL '05, pages 83–90.

[Monod and Jacob, 1961] Monod, J. and Jacob, F. (1961). General conclusions: Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb. Symp. Quant. Biol.*, 26(0):389–401.

[Mooney et al., 2006] Mooney, R. J., Bunescu, R. C., and Mooney, R. J. (2006). Subsequence kernels for relation extraction. *Adv. Neural Inf. Process. Syst.*, 18:171.

[Morgan et al., 2004] Morgan, A. A., Lynette, H., Marc, C., Yeh, A. S., and Colombe, J. B. (2004). Gene name identification and normalization using a model organism database. *J. Biomed. Inform.*, 37(6):396–410.

[Moschitti and Zanzotto, 2007] Moschitti, A. and Zanzotto, F. M. (2007). Fast and effective kernels for relational learning from texts. In *Proceedings of the 24th international conference on Machine learning*, pages 649–656. ACM.

[Müller et al., 2004] Müller, H.-M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, 2(11):e309.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1991):3–26.

[Nakamura et al., 2001] Nakamura, S., Lynch, T. J., and Finkelstein, R. R. (2001). Physical interactions between ABA response loci of arabidopsis. *Plant J.*, 26(6):627–635.

[Nakashima et al., 2006] Nakashima, K., Fujita, Y., Katsura, K., Maruyama, K., Narusaka, Y., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2006). Transcriptional regulation of ABI3- and ABA-responsive genes including RD29B and RD29A in seeds, germinating embryos, and seedlings of arabidopsis. *Plant Mol. Biol.*, 60(1):51–68.

[Narayanaswamy et al., 2003] Narayanaswamy, M., Ravikumar, K. E., and Vijay-Shanker, K. (2003). A biological named entity recognizer. *Pac. Symp. Biocomput.*, pages 427–438.

[Navigli and Ponzetto, 2010] Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

[Nédellec, 2005a] Nédellec, C. (2005a). Learning language in logic - genic interaction extraction challenge. In *Proceedings of the Learning Language in Logic 2005 Workshop at the International Conference on Machine Learning*.

[Nédellec, 2005b] Nédellec, C. (2005b). Learning language in logic-genic interaction extraction challenge. In *Proc ICML-05, Learning Language in Logic Workshop, LLL-05*, pages 31–37.

[Nédellec, 2013] Nédellec, C. (2013). Extraction et modélisation de connaissance à partir de texte – applications à la biologie. Master's thesis, Université Blaise Pascal.

[Nédellec et al., 2006] Nédellec, C., Bessières, P., Bossy, R., Kotoujansky, A., and Manine, A.-P. (2006). Annotation guidelines for machine Learning-Based named entity recognition in microbiology. In *Workshop on Data and Text Mining for Integrative Biology*, page 40. ecmlpkdd2006.org.

[Nédellec et al., 2014] Nédellec, C., Bossy, R., Valsamou, D., Ranoux, M., Golik, W., and Sourdille, P. (2014). Information extraction from bibliography for Marker-Assisted selection in wheat. In *Metadata and Semantics Research*, Communications in Computer and Information Science, pages 301–313. Springer International Publishing.

[Nédellec and Dubreucq, 2014] Nédellec, C. and Dubreucq, B. (2014). Extraire et formaliser les connaissances pour la biologie ou les bases du dialogue entre biologie et informatique.

[Nédellec et al., 2010] Nédellec, C., Golik, W., Aubin, S., and Bossy, R. (2010). Building large lexicalized ontologies from text: a use case in automatic indexing of biotechnology patents. In *Knowledge Engineering and Management by the Masses*, pages 514–523. Springer.

[Nédellec et al., 2009] Nédellec, C., Nazarenko, A., and Bossy, R. (2009). Information extraction. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, pages 663–685. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Needleman and Wunsch, 1970] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453.

[Neuhaus and Bunke, 2006] Neuhaus, M. and Bunke, H. (2006). Edit distance-based kernel functions for structural pattern classification. *Pattern Recognit.*, 39(10):1852–1863.

[Ng and Wong, 1999] Ng, S. K. and Wong, M. (1999). Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Inform. Ser. Workshop Genome Inform.*, 10:104–112.

[Ng and Cardie, 2002] Ng, V. and Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

[Nguyen et al., 2015] Nguyen, T. H., Plank, B., and Grishman, R. (2015). Semantic representations for domain adaptation: a case study on the tree kernel-based method for relation extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

*Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing*, pages 635–644.

[Nodine and Bartel, 2010] Nodine, M. D. and Bartel, D. P. (2010). MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. *Genes Dev.*, 24(23):2678–2692.

[North et al., 2010] North, H., Baud, S., Debeaujon, I., Dubos, C., Dubreucq, B., Grappin, P., Jullien, M., Lepiniec, L., Marion-Poll, A., Miquel, M., et al. (2010). Arabidopsis seed secrets unravelled after a decade of genetic and omics-driven research. *The Plant Journal*, 61(6):971–981.

[NSF, 2000] NSF (2000). Nifty 50: ARABIDOPSIS – a PLANT GENOME PROJECT. http://www.nsf.gov/od/lpa/nsf50/nsfoutreach/htm/n50_z2/pages_z3/05_pg.htm.

[NSF, 2013] NSF (2013). Arabidopsis: The model plant. http://www.nsf.gov/pubs/2002/bio0202/model.htm.

[Ogas et al., 1999] Ogas, J., Kaufmann, S., Henderson, J., and Somerville, C. (1999). PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.*, 96(24):13839–13844.

[Ogawa et al., 2007] Ogawa, T., Uchimiya, H., and Kawai-Yamada, M. (2007). Mutual regulation of arabidopsis thaliana ethylene-responsive element binding protein and a plant floral homeotic gene, APETALA2. *Ann. Bot.*, 99(2):239–244.

[Ogren, 2006] Ogren, P. V. (2006). Knowtator: A ProtÉGÉ plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume: Demonstrations*, NAACL-Demonstrations '06, pages 273–275, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Ohta et al., 2013a] Ohta, T., Pyysalo, S., Rak, R., Rowley, A., Chun, H.-W., Jung, S.-J., Choi, S.-P., Ananiadou, S., and Tsujii, J. (2013a). Overview of the pathway curation (pc) task of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 67–75, Sofia, Bulgaria. Association for Computational Linguistics.

[Ohta et al., 2013b] Ohta, T., Pyysalo, S., and Tsujii, J. (2013b). Proceedings of the BioNLP shared task 2013 workshop.

[Ohto et al., 2005] Ohto, M.-A., Fischer, R. L., Goldberg, R. B., Nakamura, K., and Harada, J. J. (2005). Control of seed mass by APETALA2. *Proc. Natl. Acad. Sci. U. S. A.*, 102(8):3123–3128.

[Oliveros et al., 2000] Oliveros, J. C., Blaschke, C., Herrero, J., Dopazo, J., and Valencia, A. (2000). Expression profiles and biological function. *Genome Inform. Ser. Workshop Genome Inform.*, 11:106–117.

[Ono et al., 2001] Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161.

[Özgür and Radev, 2009] Özgür, A. and Radev, D. R. (2009). Supervised classification for extracting biomedical events. In *Proceedings of the Workshop on BioNLP Shared Task BioNLP 09*, page 111.

[Papazian et al., 2012] Papazian, F., Bossy, R., and Nédellec, C. (2012). AlvisAE: A collaborative web text annotation editor for knowledge acquisition. In *Proceedings of the Sixth Linguistic Annotation Workshop*, LAW VI '12, pages 149–152, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Park et al., 2001] Park, J. C., Kim, H. S., and Kim, J. J. (2001). Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Pac. Symp. Biocomput.*, pages 396–407.

[Pedersen et al., 2004] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::Similarity: Measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Peng and Weselake, 2011] Peng, F. Y. and Weselake, R. J. (2011). Gene coexpression clusters and putative regulatory elements underlying seed storage reserve accumulation in arabidopsis. *BMC Genomics*, 12:286.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Perry et al., 1999] Perry, S. E., Lehti, M. D., and Fernandez, D. E. (1999). The MADS-domain protein AGAMOUS-like 15 accumulates in embryonic tissues with diverse origins. *Plant Physiol.*, 120(1):121–130.

[Pogodalla, 2009] Pogodalla, S. (2009). Corpus linguistics (lecture notes).

[Pulavarthi et al., 2000] Pulavarthi, P., Chiang, R., and Altman, R. B. (2000). Generating interactive molecular documentaries using a library of graphical actions. *Pac. Symp. Biocomput.*, pages 266–277.

[Pustejovsky et al., 2002] Pustejovsky, J., Castaño, J., Zhang, J., Kotecki, M., and Cochran, B. (2002). Robust relational parsing over biomedical literature: extracting inhibit relations. *Pac. Symp. Biocomput.*, pages 362–373.

[Pyysalo et al., 2007] Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50.

[Ramani et al., 2005] Ramani, A. K., Bunescu, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.*, 6(5):R40.

[Rao et al., 2013] Rao, D., McNamee, P., and Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing, pages 93–115. Springer Berlin Heidelberg.

[Rashotte et al., 2006] Rashotte, A. M., Mason, M. G., Hutchison, C. E., Ferreira, F. J., Schaller, G. E., and Kieber, J. J. (2006). A subset of arabidopsis AP2 transcription factors mediates cytokinin responses in concert with a two-component pathway. *Proc. Natl. Acad. Sci. U. S. A.*, 103(29):11081–11085.

[Ratkovic, 2014] Ratkovic, Z. (2014). *Predicative Analysis for Information Extraction.* PhD thesis, Université Sorbonne Nouvelle - Paris 3.

[Ratkovic et al., 2012] Ratkovic, Z., Golik, W., and Warnier, P. (2012). Event extraction of bacteria biotopes: a knowledge-intensive nlp-based approach. *BMC Bioinformatics*, 13(Suppl 11).

[Ratkovic et al., 2011] Ratkovic, Z., Golik, W., Warnier, P., Veber, P., and Nédellec, C. (2011). BioNLP 2011 task bacteria biotope: the alvis system. In *Proceedings of the BioNLP Workshop at ACL Conference*, pages 102–111.

[Ray and Craven, 2001] Ray, S. and Craven, M. (2001). Representing sentence structure in hidden markov models for information extraction. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 1273–1279. cs.columbia.edu.

[Raychaudhuri and Altman, 2003] Raychaudhuri, S. and Altman, R. B. (2003). A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, 19(3):396–401.

[Rebholz-Schuhmann et al., 2005] Rebholz-Schuhmann, D., Kirsch, H., and Couto, F. (2005). Facts from text–is text mining ready to deliver? *PLoS Biol.*, 3(2):e65.

[Reichartz et al., 2010] Reichartz, F., Korte, H., and Paass, G. (2010). Semantic relation extraction with kernels over typed dependency trees. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 773–782, New York, NY, USA. ACM.

[Resnik, 1995] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy.

[Reza et al., 2011] Reza, A., Vincent, E., and Pascale, C. (2011). Using shallow linguistic features for relation extraction in bio-medical texts. *Knowledge Management*, 3(1).

[Riedel et al., 2010] Riedel, S., Yao, L., and Mccallum, A. (2010). Modeling relations and their mentions without labeled text. *Machine Learning and Knowledge*, pages 1–16.

[Riloff, 1993] Riloff, E. (1993). Automatically constructing a dictionary for information extraction tasks. In *Association for the Advancement of Artificial Intelligence*, pages 811–816.

[Rimell and Clark, 2008] Rimell, L. and Clark, S. (2008). Adapting a lexicalized-grammar parser to contrasting domains. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 475–484.

[Rinaldi et al., 2007] Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Andronis, C., Konstandi, O., and Persidis, A. (2007). Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artif. Intell. Med.*, 39(2):127–136.

[Roberts et al., 2008] Roberts, P. M., Cohen, A. M., and Hersh, W. R. (2008). Tasks, topics and relevance judging for the TREC genomics track: five years of experience evaluating biomedical text information retrieval systems. *Inf. Retr. Boston.*, 12(1):81–97.

[Rowley and Hartley, 2008] Rowley, J. E. and Hartley, R. J. (2008). *Organizing Knowledge: An Introduction to Managing Access to Information.* Ashgate Publishing, Ltd.

[Rzhetsky et al., 2000] Rzhetsky, A., Koike, T., Kalachikov, S., Gomez, S. M., Krauthammer, M., Kaplan, S. H., Kra, P., Russo, J. J., and Friedman, C. (2000). A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics*, 16(12):1120–1128.

[Sahlgren, 2008] Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

[Santos-Mendoza et al., 2008] Santos-Mendoza, M., Dubreucq, B., Baud, S., Parcy, F., Caboche, M., and Lepiniec, L. (2008). Deciphering gene regulatory networks that control seed development and maturation in arabidopsis. *Plant J.*, 54(4):608–620.

[Saunders et al., 2013] Saunders, D. R., Bex, P. J., and Woods, R. L. (2013). Crowdsourcing a normative natural language dataset: a comparison of amazon mechanical turk and in-lab data collection. *J. Med. Internet Res.*, 15(5):e100.

[Schölkopf and Smola, 2002] Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Adaptive computation and machine learning. MIT Press.

[Schölkopf et al., 2004] Schölkopf, B., Tsuda, K., and Vert, J. P. (2004). *Kernel Methods in Computational Biology.* A Bradford book. Bradford Bks.

[Schoof et al., 2004] Schoof, H., Ernst, R., Nazarov, V., Pfeifer, L., Mewes, H., and Mayer, K. F. X. (2004). MIPS arabidopsis thaliana database (matdb): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.*, 32(suppl 1):D373–D376.

[Schoof et al., 2002] Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G., Arnold, R., Mewes, H. W., and Mayer, K. F. X. (2002). MIPS arabidopsis thaliana database (matdb): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res.*, 30(1):91–93.

[Schutz and Buitelaar, 2005] Schutz, A. and Buitelaar, P. (2005). RelExt: A tool for relation extraction from text in ontology extension. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3729 LNCS, pages 593–606.

[Segura-Bedmar et al., 2011] Segura-Bedmar, I., Martinez, P., de Pablo-Sánchez, C., Martínez, P., de Pablo-Sánchez, C., Martinez, P., and de Pablo-Sánchez, C. (2011). Using a shallow linguistic kernel for drug–drug interaction extraction. *J. Biomed. Inform.*, 44(5):789–804.

[Segura Bedmar et al., 2013] Segura Bedmar, I., Martínez, P., and Herrero Zazo, M. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013).

[Sellers, 1974] Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.*, 26(4):787–793.

[Shatkay et al., 2000] Shatkay, H., Edwards, S., Wilbur, W. J., and Boguski, M. (2000). Genes, themes and microarrays. In *Proc. Int. Conf. Intell. Syst. Mol. Biol*, volume 8, pages 317–327.

[Singh et al., 2012] Singh, S., Subramanya, A., Pereira, F., and McCallum, A. (2012). Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012-015.*

[Skounakis et al., 2003] Skounakis, M., Craven, M., and Ray, S. (2003). Hierarchical hidden markov models for information extraction. In *IJCAI*, pages 427–433.

[Smalheiser and Swanson, 1994] Smalheiser, N. R. and Swanson, D. R. (1994). Assessing a gap in the biomedical literature-magnesium-deficiency and neurologic disease. *Neurosci. Res. Commun.*, 15(1):1–9.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197.

[Söderman et al., 2000] Söderman, E. M., Brocard, I. M., Lynch, T. J., and Finkelstein, R. R. (2000). Regulation and function of the arabidopsis ABA-insensitive4 gene in seed and abscisic acid response signaling networks. *Plant Physiol.*, 124(4):1752–1765.

[Soon et al., 2001] Soon, W. M., Ng, H. T., and Daniel (2001). A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, 27(4):521–544.

[Sørensen, 1948] Sørensen, T. J. (1948). *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.* I kommission hos E. Munksgaard, København.

[Srihari and Li, 1999] Srihari, R. and Li, W. (1999). Information extraction supported question answering.

[Srinivasan and Libbus, 2004] Srinivasan, P. and Libbus, B. (2004). Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, 20 Suppl 1:i290–6.

[Stapley and Benoit, 2000] Stapley, B. J. and Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. *Pac. Symp. Biocomput.*, pages 529–540.

[Stapley et al., 2002] Stapley, B. J., Kelley, L. A., and Sternberg, M. J. E. (2002). Predicting the sub-cellular location of proteins from text using support vector machines. *Pac. Symp. Biocomput.*, pages 374–385.

[Stenetorp et al., 2012] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Stone et al., 2001] Stone, S. L., Kwong, L. W., Yee, K. M., Pelletier, J., Lepiniec, L., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (2001). LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development. *Proc. Natl. Acad. Sci. U. S. A.*, 98(20):11806–11811.

[Studer et al., 1998] Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data Knowl. Eng.*, 25(1–2):161–197.

[Su et al., 2008] Su, J., Yang, X., Hong, H., Tateisi, Y., and Tsujii, J. (2008). Coreference resolution in biomedical texts: a machine learning approach. *Ontologies and Text Mining for Life Sciences*, 8.

[Surdeanu and Tibshirani, 2012] Surdeanu, M. and Tibshirani, J. (2012). Multi-instance multi-label learning for relation extraction. *Proceedings of the 2012 ...*.

[Suzuki et al., 2007] Suzuki, M., Wang, H. H.-Y., and McCarty, D. R. (2007). Repression of the LEAFY COTYLEDON 1/b3 regulatory network in plant embryo development by VP1/ABSCISIC ACID INSENSITIVE 3-LIKE B3 genes. *Plant Physiol.*, 143(2):902–911.

[Swanson, 1988] Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspect. Biol. Med.*, 31(4):526–557.

[Szakonyi et al., 2015] Szakonyi, D., Van Landeghem, S., Baerenfaller, K., Baeyens, L., Blomme, J., Casanova-Sáez, R., De Bodt, S., Esteve-Bruna, D., Fiorani, F., Gonzalez, N., Grønlund, J., Immink, R. G. H., Jover-Gil, S., Kuwabara, A., Muñoz Nortes, T., van Dijk, A. D. J., Wilson-Sánchez, D., Buchanan-Wollaston, V., Angenent, G. C., Van de Peer, Y., Inzé, D., Micol, J. L., Gruissem, W., Walsh, S., and Hilson, P. (2015). The KnownLeaf literature curation system captures knowledge about arabidopsis leaf growth and development and facilitates integrated data mining. *Current Plant Biology*, pages 1–11.

[TAIR, 2015] TAIR (2015). TAIR - about arabidopsis. https://www.arabidopsis.org/portals/education/aboutarabidopsis.jsp.

[Tamames et al., 1998] Tamames, J., Ouzounis, C., Casari, G., Sander, C., and Valencia, A. (1998). EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, 14(6):542–543.

[Tari, 2013] Tari, L. (2013). Knowledge inference. In *Encyclopedia of Systems Biology*, pages 1074–1078. Springer New York.

[Thomas et al., 2011] Thomas, P., Neves, M., Solt, I., Tikk, D., and Leser, U. (2011). Relation extraction for drug-drug interactions using ensemble learning. *Training*, pages 1–8.

[Tikk et al., 2010] Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., and Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput. Biol.*, 6:e1000837.

[To et al., 2006] To, A., Valon, C., Savino, G., Guilleminot, J., Devic, M., Giraudat, J., and Parcy, F. (2006). A network of local and redundant gene regulation governs arabidopsis seed maturation. *Plant Cell*, 18(7):1642–1651.

[Tsatsaronis et al., 2015] Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artiéres, T., Ngomo, A.-C. N., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., and Paliouras, G. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

[Tsuchiya et al., 2004] Tsuchiya, Y., Nambara, E., Naito, S., and McCourt, P. (2004). The FUS3 transcription factor functions through the epidermal regulator TTG1 during embryogenesis in arabidopsis. *Plant J.*, 37(1):73–81.

[Tsuda, 1999] Tsuda, K. (1999). Support vector classifier with asymmetric kernel functions. In *in European Symposium on Artificial Neural Networks (ESANN)*.

[Tsuruoka et al., 2005] Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics*, pages 382–392. Springer.

[Tsuruoka and Tsujii, 2004] Tsuruoka, Y. and Tsujii, J. (2004). Improving the performance of dictionary-based approaches in protein name recognition. *J. Biomed. Inform.*, 37(6):461–470.

[Uzuner et al., 2011] Uzuner, Ö., South, B. R., Uzuner, O., South, B. R., Shen, S., and Duvall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.*, 18(5):552–556.

[Valsamou, 2013] Valsamou, D. (2013). Extraction d'information pour la reconstruction des réseaux de régulation biologiques impliquées dans le développement de la graine chez a. thaliana.

[Valsamou, 2015] Valsamou, D. (2015). Apprentissage automatique pour l'extraction de réseaux de régulation géniques à partir d'articles.

[Valsamou and Nedellec, 2012] Valsamou, D. and Nedellec, C. (2012). Relation extraction from biological text.

[Van Auken et al., 2012] Van Auken, K., Fey, P., Berardini, T. Z., Dodson, R., Cooper, L., Li, D., Chan, J., Li, Y., Basu, S., Muller, H.-M., Chisholm, R., Huala, E., and Sternberg, P. W. (2012). Text mining in the biocuration workflow: applications for literature curation at WormBase, dictybase and TAIR. *Database*, 2012:bas040.

[Van Landeghem et al., 2013] Van Landeghem, S., De Bodt, S., Drebert, Z. J., Inzé, D., and Van de Peer, Y. (2013). The potential of text mining in data integration and network biology for plant research: a case study on arabidopsis. *Plant Cell*, 25(3):794–807.

[van Mulligen et al., 2012] van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J. A., and Furlong, L. I. (2012). The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *J. Biomed. Inform.*, 45(5):879–884.

[Veber et al., 2011] Veber, P., Bessières, P., Nédellec, C., Bossy, R., Golik, W., Jourde, J., Warnier, P., and Ratkovic, Z. (2011). INRA's quarterly report internal deliverable CTC-16, quaero program, CTC project. Technical Report 16, Institut National de la Recherche Agronomique.

[Vicient et al., 2000] Vicient, C. M., Bies-Etheve, N., and Delseny, M. (2000). Changes in gene expression in the leafy cotyledon1 (lec1) and fusca3 (fus3) mutants of arabidopsis thaliana L. *J. Exp. Bot.*, 51(347):995–1003.

[Voormann and Gut, 2008] Voormann, H. and Gut, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251.

[Wallis, 2007] Wallis, S. (2007). Annotation, retrieval and experimentation. *Annotating Variation and Change. Helsinki: Varieng,[University of Helsinki].*

[Wang et al., 2004] Wang, H., Caruso, L. V., Downie, A. B., and Perry, S. E. (2004). The embryo MADS domain protein AGAMOUS-Like 15 directly regulates expression of a gene encoding an enzyme involved in gibberellin metabolism. *Plant Cell*, 16(5):1206–1219.

[Weeber et al., 2003] Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L. T. W., Aronson, A. R., and Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J. Am. Med. Inform. Assoc.*, 10(3):252–259.

[Weijers and Jürgens, 2005] Weijers, D. and Jürgens, G. (2005). Auxin and embryo axis formation: the ends in sight? *Curr. Opin. Plant Biol.*, 8(1):32–37.

[Western et al., 2004] Western, T. L., Young, D. S., Dean, G. H., Tan, W. L., Samuels, A. L., and Haughn, G. W. (2004). MUCILAGE-MODIFIED4 encodes a putative pectin biosynthetic enzyme developmentally regulated by APETALA2, TRANSPARENT TESTA GLABRA1, and GLABRA2 in the arabidopsis seed coat. *Plant Physiol.*, 134(1):296–306.

[Widlöcher and Mathet, 2012] Widlöcher, A. and Mathet, Y. (2012). The glozz platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, DocEng '12, pages 171–180, New York, NY, USA. ACM.

[Willmann et al., 2011] Willmann, M. R., Mehalick, A. J., Packer, R. L., and Jenik, P. D. (2011). MicroRNAs regulate the timing of embryo maturation in arabidopsis. *Plant Physiol.*, 155(4):1871–1884.

[Winkler, 1990] Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage.

[Winter et al., 2007] Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G. V., and Provart, N. J. (2007). An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. *PloS one*, 2(8):e718.

[Wu and Palmer, 1994] Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Würschum et al., 2006] Würschum, T., Gross-Hardt, R., and Laux, T. (2006). APETALA2 regulates the stem cell niche in the arabidopsis shoot meristem. *Plant Cell*, 18(2):295–307.

[Yakushiji et al., 2006] Yakushiji, A., Miyao, Y., Ohta, T., Tateisi, Y., and Tsujii, J. (2006). Automatic construction of predicate-argument structure patterns for biomedical information extraction. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing EMNLP 06*, (July):284.

[Yakushiji et al., 2001] Yakushiji, A., Tateisi, Y., Miyao, Y., and Tsujii, J. (2001). Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput.*, 419:408–419.

[Yamamoto et al., 2003] Yamamoto, K., Kaoru, Y., Taku, K., Akihiko, K., and Yuji, M. (2003). Protein name tagging for biomedical annotation in text. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine -*.

[Yandell and Majoros, 2002] Yandell, M. D. and Majoros, W. H. (2002). Genomics and natural language processing. *Nat. Rev. Genet.*, 3(8):601–610.

[Yangarber et al., 2000] Yangarber, R., Roman, Y., Ralph, G., Pasi, T., and Silja, H. (2000). Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th conference on Computational linguistics -*.

[Yeganova et al., 2004] Yeganova, L., Smith, L., and Wilbur, W. J. (2004). Identification of related gene/protein names based on an HMM of name variations. *Comput. Biol. Chem.*, 28(2):97–107.

[Yeh et al., 2002] Yeh, A., Hirschman, L., and Morgan, A. (2002). Background and overview for KDD cup 2002 task 1: Information extraction from biomedical articles. *SIGKDD Explor. Newsl.*, 4(2):87–89.

[Yeh et al., 2003] Yeh, A. S., Hirschman, L., and Morgan, A. A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD challenge cup. *Bioinformatics*, 19(Suppl 1):i331–i339.

[Yilmaz et al., 2011] Yilmaz, A., Mejia-Guerra, M. K., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). Agris: the arabidopsis gene regulatory information server, an update. *Nucleic acids research*, 39(suppl 1):D1118–D1122.

[Yu et al., 2012] Yu, D., Jiang, L., Gong, H., and Liu, C.-M. (2012). EMBRYONIC FACTOR 19 encodes a pentatricopeptide repeat protein that is essential for the initiation of zygotic embryogenesis in arabidopsis. *J. Integr. Plant Biol.*, 54(1):55–64.

[Zelenko et al., 2003] Zelenko, D., Aone, C., and Richardella, A. (2003). Kernel methods for relation extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 3(6):1083–1106.

[Zhang et al., 2006] Zhang, M., Zhang, J., Su, J., and Zhou, G. (2006). A composite kernel to extract relations between entities with both flat and structured features. *... and the 44th annual meeting of the ...*, (July):825–832.

[Zhang et al., 2008] Zhang, M., Zhou, G., and Aw, A. (2008). Exploring syntactic structured features over parse trees for relation extraction using kernel methods. *Inf. Process. Manag.*, 44(2):687–701.

[Zhang et al., 2005] Zhang, X., Garreton, V., and Chua, N.-H. (2005). The AIP2 E3 ligase acts as a novel negative regulator of ABA signaling by promoting ABI3 degradation. *Genes Dev.*, 19(13):1532–1543.

[Zhao and Grishman, 2005] Zhao, S. and Grishman, R. (2005). Extracting relations with integrated information using kernel methods. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (June):419–426.

[Zhao et al., 2004] Zhao, S., Meyers, A., and Grishman, R. (2004). Discriminative slot detection using kernel methods. In *Proceedings of the 20th international conference on Computational Linguistics*, page 757. Association for Computational Linguistics.

[Zhou et al., 2004] Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C. (2004). Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.

[Zhu and Perry, 2005] Zhu, C. and Perry, S. E. (2005). Control of expression and autoregulation of AGL15, a member of the MADS-box family. *Plant J.*, 41(4):583–594.

[Zweigenbaum et al., 2007] Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Brief. Bioinform.*, 8(5):358–375.

# Synthèse

Ce travail propose l'Extraction d'Information (EI) comme une approche efficace pour la production de l'information structuree sur la biologie. Malgré l'abondance et la couverture de la littérature scientifique en biologie, l'information contenue n'est pas structurée et directement utilisable pour la modélisation des organismes biologiques. L'approche de transformation de l'information disponible dans les articles scientifiques en donnée structurée et utilisable est décrite dans cette thèse en presentant une tache complete d'EI sur un organisme modele, *Arabidopsis thaliana*. Deux axes principaux sont définis pour accomplir cette tâche : le premier concerne les données et le modèle de connaissances, plus complexe que la majorité ceux de la majorité des tâches d'EI dans l'état de l'art, et le deuxième qui concerne l'approche algorithmique mise en place pour l'EI. Un tel systeme d'EI se charge d'extraire les parties de texte les plus significatives et d'identifier leurs relations semantiques, et dans leur majorité ces systèmes sont basés sur l'apprentissage automatique et plus précisément l'apprentissage supervisé.

Sur le volet des données et du modèle, le travail a été réalisé en collaboration avec des experts biologistes sur la plante *A. thaliana*. Afin de formaliser la connaissance nécessaire pour bien décrire le domaine du développement de la graine de *A. thaliana*, un modele de connaissance a ete concu définissant toutes les entites essentielles ainsi que leurs relations. Ce modèle est une représentation de l'information qui peut etre directement utilisee par des algorithmes. Ce modele reconcilie les besoins d'avoir un modele assez complexe pour bien decrire le domaine, et d'avoir assez de generalite pour pouvoir utiliser des methodes d'apprentissage automatique. Néanmoins, il n'est pas limité à l'usage pour l'EI, et il peut servir pour d'autres applications, telles que la construction d'ontologies.

Les experts ont également annoté un ensemble d'articles scientifiques du domaine en utilisant les entités et les relations du modèle. Cette phase d'annotation a servi comme outil d'amélioration et confirmation du modèle. Le résultat de ce travail d'annotation est le corpus nécessaire pour l'entraînement de l'apprentissage automatique utilisé dans ce travail, mais il est également rendu disponible publiquement. Le modele et le corpus annote sont les premiers proposes pour le developpement de la graine, et parmi les rares pour *A. thaliana*, malgre son importance biologique. Suite aux travaux décrits dans cette thèse, ils ont été utilisés pour d'autres projets (*p.e.* OpenMinteD) mais également dans le cadre de challenges internationaux (*p.e.* SeeDev dans BioNLP ST 16).

En parallèle avec les travaux sur le modèle et les données AlvisRE, une approche d'extraction de relations, a egalement ete elaboree et developpee. Dans les systèmes d'EI, l'extraction de relations intervient une fois les entités reconnues, et l'extracteur de relations cherche à détecter les cas où le texte mentionne une relation entre elles, puis identifier précisément de quel type de relation du modèle il s'agit. L'approche AlvisRE est basee sur la similarite textuelle et utilise a la fois des informations lexiques, syntactiques et semantiques, lui permettant de profiter de niveaux d'abstraction variés pour mieux classifier les relations extraites selon le modèle. Dans les experiences realisees, AlvisRE donne des resultats qui sont equivalents et parfois superieurs a l'etat de l'art. En plus, AlvisRE a l'avantage de la modularite et adaptabilite à de nouveaux corpus et cas d'usage, grâce au fait que le système peut profiter des plusieurs niveaux d'abstraction et, notamment, d' informations semantiques produites automatiquement. Ce dernier caracteristique permet d'attendre des performances equivalentes dans d'autres domaines, lors des expériences réalisés au cadre de cette thèse.

université
PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies
de l'information
et de la communication (STIC)

**Titre :** Extraction d'Information pour les réseaux de régulation de la graine chez *Arabidopsis Thaliana.*

**Mots clefs :** Extraction d'Information, Extraction de Relations, Traitement Automatique du Langage, TAL, BioNLP, Bioinformatique

**Résumé :** Ce travail propose l'Extraction d'Information (EI) comme une approche efficace pour la production de l'information structurée, utilisable sur la biologie, en présentant une tâche complète d'EI sur un organisme modèle, *Arabidopsis thaliana.* Un système d'EI se charge d'extraire les parties de texte les plus significatives et d'identifier leurs relations sémantiques.
En collaboration avec des experts biologistes sur la plante *A. Thaliana* un modèle de connaissance a été conçu danns l'objectif de formaliser la connaissance nécessaire pour bien décrire le domaine du développement de la graine. Ce modèle contient toutes les entités et relations les connectant qui sont essentielles et peut être directement utilisé par des algorithmes. En parallèle ce modèle a été testé et appliqué sur un ensemble d'articles scientifiques du domaine, le corpus nécessaire pour l'entraînement de l'apprentissage automatique, annoté en utilisant les entités et relations du modèle. Le modèle et le corpus annoté sont les premiers proposés pour le développement de la graine, et parmi les rares pour *A. Thaliana*, malgré son importance biologique. Ce modèle réconcilie les besoins d'avoir un modèle assez complexe pour bien décrire le domaine, et d'avoir assez de généralité pour pouvoir utiliser des méthodes d'apprentissage automatique.
Une approche d'extraction de relations (AlvisRE) a également été élaborée et développée. L'approche AlvisRE est basée sur la similarité textuelle et utilise à la fois des informations lexiques, syntactiques et sémantiques. Dans les expériences réalisées, AlvisRE donne des résultats qui sont équivalents et parfois supérieurs à l'état de l'art. En plus, AlvisRE a l'avantage de la modularité et adaptabilité en utilisant des informations sémantiques produites automatiquement. Ce dernier caractéristique permet d'attendre des performances équivalentes dans d'autres domaines.

**Title :** Information Extraction for the Seed Development Regulatory Networks of *Arabidopsis Thaliana.*

**Keywords :** Information Extraction, Relation Extraction, Natural Language Processing, NLP, BioNLP, Bioinformatics.

**Abstract :** This work proposes Information Extraction (IE) as an efficient approach for producing structured, usable information on biology, by presenting a complete IE task on a model biological organism, *Arabidopsis thaliana.* Information Extraction is the process of extracting meaningful parts of text and identifying their semantic relations.
In collaboration with experts on the plant *A. Thaliana*, a knowledge model was conceived, with the goal of providing a formal representation of the knowledge that is necessary to sufficiently describe the domain of grain development. This model contains all the entities and the relations between them which are essential and it can directly be used by algorithms. In parallel, this model was tested and applied on a set of scientific articles of the domain. These documents constitute the necessary corpus for training machine learning algorithms, annotated by experts using the entities and relations of the model. This corpus and this model are the first available for grain development and among very few on *A. Thaliana*, despite the latter's importance in biology. This model manages to answer both needs of being complex enough to describe the domain well, and of having enough generalization for machine learning.
A relation extraction approach (AlvisRE) was also elaborated and developed. AlvisRE's approach is based on textual similarity and it uses all types of information available: lexical, syntactic and semantic. In the tests conducted, AlvisRE had results that are equivalent or sometimes better than the state of the art. Additionally, AlvisRE has the advantage of being modular and adaptive by using semantic information that was produced automatically. This last feature allows me to expect similar performance in other domains.

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France