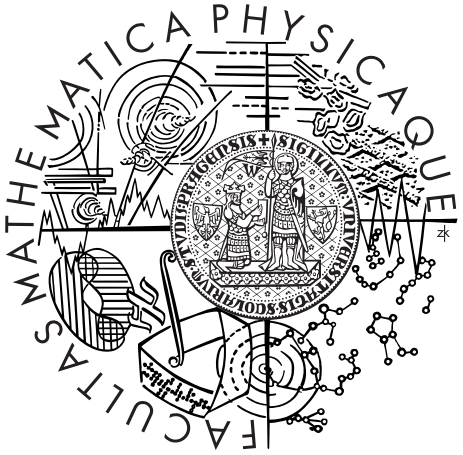
Charles University in Prague

Faculty of Mathematics and Physics

MASTER THESIS



Jan Václ

Tracing Salience in Texts

Institute of Formal and Applied Linguistics

Supervisor of the master thesis: Mgr. Barbora Vidová Hladká, Ph.D.

Study programme: Informatics

Specialization: Mathematical Linguistics

Prague 2015

I would like to thank Mgr. Barbora Vidová Hladká for her constant help and assistance with this thesis, prof. Eva Hajičová for her great inspiration, and finally my parents for their admirable support and patience.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In ..... date ....

Název práce: Sledování aktivovanosti objektů v textech

Autor: Jan Václ

Vedoucí diplomové práce: Mgr. Barbora Vidová Hladká, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: V kontextu analýzy diskurzu stupeň aktivovanosti (salience) modeluje aktuální míru zapojenosti odkazovaných objektů a její vývoj v průběhu textu. Algoritmus pro určování aktivovanosti a vizualizaci jejího průběhu již byl navržen a otestován na malém vzorku dat. Tato práce reprodukuje výsledky algoritmu ve větším měřítku pomocí dat z Pražského závislostního korpusu 3.0. Výsledky jsou pak zpracovány do přístupného tvaru a je provedena jejich analýza jak pomocí vizuálního výstupu, tak i výstupů kvantitativních. Přitom jsou zohledněny dva základní stavební kameny aktivovanosti; koreferenční vztahy a informační struktura věty. Práce se také zabývá možností modelování aktivovanosti pomocí strojového učení za použití algoritmů rozhodovacích stromů a náhodného lesa. V závěrečné části je proveden experiment zkoumající možné využití informace o aktivovanosti v některé z úloh strojového učení při zpracování přirozeného jazyka na příkladě shlukování dokumentů.

Klíčová slova: aktivovanost, salience, koreference, TFA, strojové učení

Title: Tracing Salience in Texts

Author: Jan Václ

Supervisor: Mgr. Barbora Vidová Hladká, Ph.D., Institute of Formal and Applied Linguistics

Abstract: The notion of salience in the discourse analysis models how the activation of referred objects evolves in the flow of text. The salience algorithm was already defined and tested briefly in an earlier research, we present a reproduction of its results in a larger scale using data from the Prague Dependency Treebank 3.0. The results are then collected into an accessible shape and analyzed both in their visual and quantitative form in the context of the two main resources of the salience – coreference relations and topic-focus articulation. Furthermore, a possibility of modeling the salience degree by a machine learning algorithm (decision trees and random forest) is examined. Finally, attempts are made with using the salience information in the machine learning NLP task of document clustering visualization.

Keywords: salience, coreference, TFA, machine learning

Contents

Introduction

Motivation

Discourse can be viewed as a sequence of sentences referring to a set of real-world objects. By modeling the dynamic appearance of these references throughout the text, one can acquire a new knowledge about the structure of the text, the importance of these objects in relation to the text, or even the nature of the text. This knowledge can be subsequently used for further investigation in the field of discourse analysis (e.g. for comparison of the discourse dynamics between two different languages), as well as enhancing the efficiency of an NLP application working with the whole bodies of texts (text segmentation, topic modeling, document clustering, information retrieval).

Goals and Contents

There are two main goals of this work. The first one is to investigate more deeply the notion of *salience* as it is defined by Hajičová et al. (2006). This includes reproducing the experiment described there on a larger amount of data (using the recently available Prague Dependency Treebank 3.0), generating the results in a human-examinable form, and analyzing them especially from the quantitative point of view.

The second goal is to examine the salience and its usefulness as an additional feature for an NLP application. Firstly, the possibility of an automatic estimation and its success rate needs to be explored, to state whether the salience value can be made available in such application. Secondly, an exemplar NLP application can be chosen and examined using the salience feature. Since the type of information the salience brings is closely associated with the topic or theme of a document, an attempt to visualize a thematic document clustering will be made.

Content Overview

The chapters in this thesis are ordered roughly from the theory to experiments and results, to provide first the context necessary for understanding the practical parts. In Chapter 1, the reader is introduced to the research context of our topic, being acknowledged with the related works in the fields we reach into. References

to the historical research for the *salience* notion are presented together with works concerning its main building blocks, *coreference* and *topic-focus articulation*.

Chapter 2 presents a necessary overview of both the linguistic theories behind this work and the algorithmic foundations of the machine learning methods used in the following chapters. Although this overview is not intended to be exhaustive and too detailed, it should provide the reader with the knowledge needed to understand this work and the presented results, along with directions to further reading if he is be more interested in any of the subjects.

Data and tools used during the experiments and other parts of the work are enlisted and described in Chapter 3, each with a brief information of how and when they contributed.

Perhaps the main part of this work is described in Chapter 4, where the results of the automatic salience analysis on the larger amount of data are presented, both quantified and visualized. These results are preceded by more general statistics of the data in question, providing the necessary context for a better interpretation.

In Chapter 5, series of experiments were conducted for exploring the possibility of predicting the salience degrees automatically, using a statistic defined in the previous chapter. Decision trees and random forest models are used in this experiments, using various features extracted from surface, morphological and syntactic layer of annotation. Aside from the evaluation results itself, the machine learning models are used also for a further understanding of the features' influence on the salience expression form.

Chapter 6 describes two experiments performed to assess the possible contribution of *salience* information as a feature to a machine learning application. Both present an attempt for a simple document clustering based on the importance of words – comparing the difference of influence of a simple word-count statistic and a salience-based feature used for calculating this importance.

Contents of the enclosed CD-ROM are described in the Appendix, along with a brief information of each piece of the data and directions on how to approach them. Both the scripts and the results contained on the CD-ROM are an integral part of this thesis and represents an important amount of work done within its scope.

1. Related Work

1.1 Discourse Dynamics and Sentence Structure

Several approaches to the analysis of a discourse dynamics with respect to a sentence structure can be found among the linguistic theories. Mainly, they attempt to capture the impact of sentence-level expressions on the flow of discourse and its topics. Most of these theories are based on distinguishing two main semantic types of information in the sentence: *given* vs. *new* (although their terminology varies, often without significant differences in the definitions).

Hajičová (2013) mentions an interesting approach to relating the sentence structure with a dynamicity of the discourse structure, given by Prince (1981); a three-level hierarchy of *givenness* of an information (contrasting the given-new) between speaker and hearer is presented there. Each level refers to a different understanding of givenness in the works of previous researchers:

1. givenness as a predictability/recoverability, as defined by Kuno (1972) and Halliday (1967) (althour their definitions slightly differ),
2. givenness in the sense of saliency, relating to the assumption of the hearer's consciousness, referring to Chafe (1976),
3. givenness in the relation to a state of a "shared knowledge" according to Haviland and Clark (1974), focusing on what the hearer "already knows and accepts to be true" vs. what the hearer "does not yet know".

Prince then continues with defining a more fine-grained familiarity scale on discourse entities, working also with the hearer's ability to infer or link the newly mentioned entities. Another "givenness hierarchy" is presented by Gundel et al. (1993) focusing on success of nominal expression referents.

Another well-known approach of modeling discourse dynamics in terms of sentence structure is the *centering* theory introduced by Joshi and Weinstein (1981) and further refined by Grosz et al. (1995), based on the local attentional states of speaker and hearer. It operates with a forward and backward looking centers of sentences and defines four types of sentence transitions by the relations of their centers. One of the characteristic features of this theory is ranking of the centers according to a language-specific parametrization.

An entity-grid model is presented by Barzilay and Lapata (2008), where each entity appearing in a text (based on a coreference relations) is assigned a column

in a grid, each sentence corresponds to a row in this grid. The cells are then filled with syntactic roles of the entities in the corresponding sentence, recording also the transitions between those sentences. It should be noted that this approach, among all the mentioned so far, is the most computationally oriented. Distributional information about the entities are extracted naturally from the entity-grid as well, forming the parameter of *salience* as a discourse prominence. However, our understanding of this notion is slightly different, agreeing with Hajičová (2013) that it should be understood in a more complex way, and that neither frequency nor the length of the referential chain is a sufficient measure of salience.

Even more application-oriented approach is presented by Sauper et al. (2010), building a statistical-based model of content structure for using it in a text analysis. This model combines hidden Markov models and conditional random fields, employing the expectation-maximization technique for finding their parameters.

1.2 Salience

Our approach directly follows the notion of *salience* first mentioned and described by Hajičová and Vrbová (1982), revisited by Hajičová (2003) and further refined and tested by Hajičová et al. (2006). This notion relates the dynamicity of the discourse with the information structure of its individual sentences, working with activation of the elements of knowledge shared between the speaker and the hearer.

In contrast to most of the works mentioned above, this approach postulates a continuous scale of the activation feature. Being defined in this continuous and relative sense, modeling and visualizing its dynamics comes forward naturally. As an indirect consequence, the theory itself does not suggest (nor enforce) any names for the particular salience levels. Although this might be confusing from the linguistic point of view, it may not necessary be a difficulty for a computational or machine learning approach.

2. Theory

The theory of *salience* will be introduced in Section 2.3, but first, one has to understand two main resources standing behind this notion: coreference and topic-focus articulation.

2.1 Coreference

Coreference is a concept describing a relation of two or more expressions in a text referring to the same real-world object. These expressions are called *referents*.

The key approach to coreference in this work is that the groups of coreferents join together to form a *coreference chain*. When speaking about two neighbouring members of the coreference chain and their relation, the first one is often called the *antecedent* and the second one the *anaphor*, with respect to the order of their occurence in the text. These terms describes the most typical form of the coreference called *anaphora*, when the first expression is the more specific one and the second one relates to the first one – when visualizing the coreference relations, this is often denoted by an arrow directed from the second one to the first one. The reverse case, called *cataphora* is also possible; however, the terminology differs here, the "target" of the relation is usually denoted as *cataphor* and it is now preceded by the "source", which is often called the *postcedent*.

The distinction between anaphora and cataphora is illustrated by the simple examples (1) in Czech (constructed based on our language experience). The English translations are as literal as possible to retain the structure of the original sentence. The coreference pairs in both cases are highlighted and subscripted.

(1) a. **Krabice$_{\mathbf{a}}$** byla tak těžká, že **ji$_{\mathbf{a}}$** Petr raději nechal za dveřmi.
(***The-box**$_{a}$ was so heavy that **it-OBJ**$_{a}$ Peter-SUBJ rather left behind the-door.*)

b. Ačkoliv **ho$_{\mathbf{c}}$** nikdo nezval, **Martin$_{\mathbf{c}}$** se-objevil na každém večírku.
(*Although **him**$_{c}$ no-one invited, **Martin**$_{c}$ showed-up on every party.*)

2.1.1 Grammatical and Textual Coreference

According to the approach to coreference captured in the Prague dependency treebanks[1] and described e.g. by Kučová and Hajičová (2004) (with its extension by Nedoluzhko (2011)), we distinguish two types of coreference relations in this work, *grammatical* and *textual*. The grammatical coreference in this approach is such a kind of coreference in which it is possible to pinpoint the coreferred expression on the basis of grammatical rules; it may involve a verb of control, reflexive pronouns, verbal complements, reciprocity and relative pronouns. On the other hand, the textual coreference is not realised by grammatical means alone, but also via context. The former type of coreference usually occurs with both the involved coreferents within one sentence, while the latter often cross the sentence boundaries.

2.1.2 Bridging Anaphora

The term *bridging anaphora*, also sometimes denoted as *associative anaphora*, is used in this work in correspondence to its annotation in the Prague Dependency Treebank, described in detail by Nedoluzhko (2011). The term describes an anaphoric relation where the anaphor is not directly coreferential to the antecedent, but an indirect connection is implied. This connection can be identified by the reader often using a real-world knowledge and a cognitive process, sometimes also based on the context. As it is shown in (2) (taken from Nedoluzhko (2011)), some knowledge of semantic structures of the mentioned object has to be employed to recognize the relationship between *"classroom"* and *children*.

(2) **Učitel** vešel do **třídy**. Děti (se) okamžitě přestaly bavit.
(***Teacher*** *entered (to)* ***the-classroom***. ***Children*** *instantly stopped talking.*)

Within the notion of *bridging anaphora*, more specific subtypes of relations are distinguished, corresponding to the semantic relation of the two referred objects. Based on a rigorous research and analysis of the impact of the inter-annotator agreement, Nedoluzhko (2011) settles for the following six subtypes for the Prague Dependency Treebank annotation task:

1. part-whole relation (asymmetric, with both possible directions)
   – e.g. *"room"-"ceiling"*, *"finger"-"hand"*

2. set-subset relation (asymmetric, with both possible directions)
   – e.g. *"drinks"-"beer"*, *"drinks"-"soda"*

[1]For details on the Prague Dependency Treebank family, see 3.1

3. functional relation (asymmetric, with both possible directions)
   – e.g. *"coach"-"team"*, *"company"-"director"*

4. semantic or pragmatic contrast (symmetric), depends heavily on the context
   – e.g. *"**Last year** we went abroad on holiday, but **this summer** we are staying at home."*

5. non-coreferential anaphoric relation (symmetric)
   – e.g. *"**Love**? What does **the word** even mean?"*

6. other – intended for collecting specific types of relations, possibly detachable into their own category in the future: family membership, place-inhabitant, author-piece, possession-owner etc.

Although some of the bridging relations are inherently asymmetric, the members of the anaphoric chain are considered to be equivalent. Thus, we can actually speak of *chains*, with each member referring to the directly previous one.

A simple example illustrating the coreference and bridging anaphora notions is given in Figure 2.1.[2] It depicts these relations among some words of the following two sentences (neighboring in the original text):

(3) Olympijský vítěz v desetiboji Robert Změlík se v minulých dnech nastěhoval se svou přítelkyní Andreou Sollárovou do nového bytu na sídlišti v Praze-Řepích. Zítra (on) vyrazí do francouzského střediska ve Font Romeu k závěrečné přípravě na mistrovství světa ve Stuttgartu.

(*The Olympic winner Robert Změlík with his girlfriend have recently moved to a new flat in a housing estate in Prague-Řepy. Tomorrow, (he) will depart to a French resort of Font Romeo to a final training before the World Championship in Stuttgart.*)

The coreference and bridging relations are marked by the colored arrows. There is a typical case of grammatical coreference (brown arrow) in the first sentence of the example: the pronoun "svou" (*"his"*) referring to the subject NP rooted in the surname Změlík. In the second sentence, the same entity is referenced again, this time by the pronoun *"he"* (zero pronoun in the Czech original, thus marked technically by *#PersPron* in this represenation). However, since this link reaches out to another sentence, it is now identified as a textual coreference (marked by the navy blue arrow). Finally, there is a part-whole bridging relation (turquoise arrow) between the resort name "Romeo" and the

[2]Both the visualization and the actual data are from the Prague Dependency Treebank – for details see further in Chapter 3

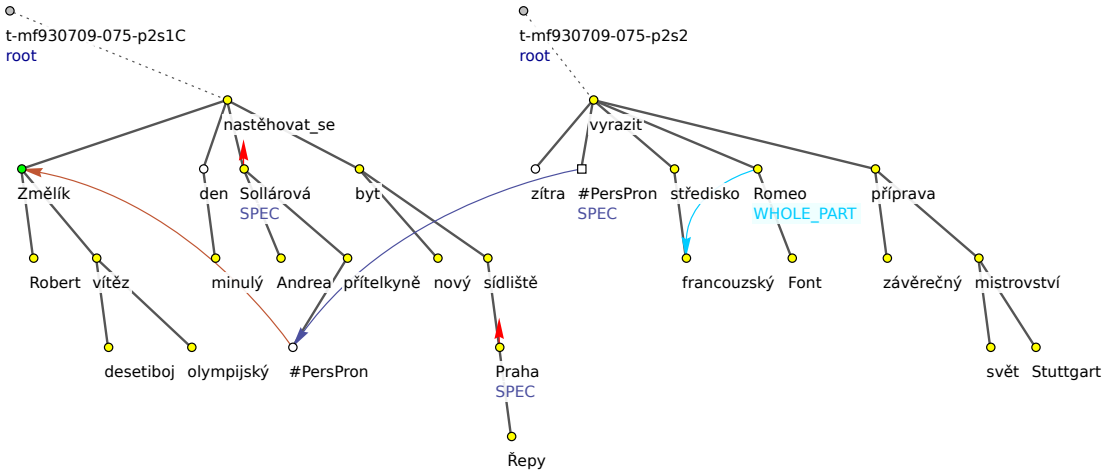


Figure 2.1: Example of visualized coreference and bridging relations in two neighboring sentences.

word "francouzský" (*"French"*). This link captures the idea that the resort as a place is "something French", i.e. part of a certain bigger complex which we denote by this adjective. There are two more red arrows, representing a textual coreference with targets outside the scope of these two sentences.

2.2 Topic-Focus Articulation

Information structure of a sentence is an important aspect of the sentence meaning, especially in the perspective of a discourse analysis. Our understanding of the sentence information structure is directly based on the Functional Generative Description framework (FGD), i.e. the approach of the Prague School of Linguistics. An insightful survey of this approach can be found in Hajičová (1993), for more detailed treatment see e.g. Sgall et al. (1986).

The key notion in this approach is the *topic-focus articulation* (or TFA), a partitioning of the sentence into two segments each with different communicational function.[3] In the *topic* part of the sentence, the speaker mentions "what he is talking about", while the *focus* part contains new information about the topic, i.e. "what he wants to say about it". The dichotomy links the semantic structure of a sentence with the structure of discourse in its context, and is usually found to be also anchored in the syntactic structure of the sentence. As described by Hajičová (1993): *"Natural languages use various surface means to convey this*

[3]This dichotomy is sometimes described also as *theme/rheme*, *topic/comment* or *presupposition/focus* by more traditional theories and also by similar contemporary approaches. However, the main distinguishing principles rarely differs.

distinction: word order plays the main role in inflectional languages, specific morphemes are present in several languages of Eastern Asia, e.g. in Japanese, and intonation seems to be important everywhere, especially in the analytic languages of Western Europe; German combines in various respects the properties of the latter with these of inflectional languages."

An example sentence in Czech (from Hajičová et al. (2005)) is shown in (4) to illustrate the topic-focus segmentation.

(4) V noci ze soboty na neděli skončil ve vojenském prostoru Ralsko sjezd majorů.
(*At night from Saturday to Sunday ended in military area Ralsko meeting(Nom.) of-majors.*)
Topic: v noci ze soboty na neděli (*at night from Saturday to Sunday*)
Focus: skončil ve vojenském prostoru Ralsko sjezd majorů (*ended in military area Ralsko meeting(Nom.) of-majors*)

As stated by Hajičová et al. (2005) and following the FGD approach, the semantic basis of the articulation of the sentence in to Topic and Focus is the relation of contextual boundness: a prototypical declarative sentence asserts that its Focus holds (or does not hold) about its Topic. Within both Topic and Focus, an opposition of contextually bound and non-bound nodes is distinguished, which is understood as a grammatically patterned opposition, rather than in the literal sense of the term. Within the contextually bound elements of the sentence, a difference is made between contrastive and non-contrastive bound elements.

Following the theoretical assumptions of FGD, TFA is captured in the tectogrammatical annotation of the Prague Dependency Treebank[4] by the TFA attribute, which may obtain one of the three values:

- t: a non-contrastive contextually bound node,
- c: a contrastive contextually bound node,
- f: a contextually non-bound node.

Returning to the relation of the two different views, the semantic view represented by the contextual boundness and non-boundness serves as a basis for inferring the syntactic, surface-form Topic/Focus dichotomy and possible segmentation of a sentence. In this direction, a heuristic procedure was proposed by Sgall et al. (1986) to identify the sentence bipartition of Topic/Focus based on the distinction of contextually bound and non-bound items.

[4]For details on the treebank, see 3.1

2.3 Salience

The flow of a discourse can be viewed as a sequence of sentences, each with its own information structure and most of them referring to some real-world objects. In different parts of the discourse, some of these objects are referred to more often than the others and vice versa. The notion of *salience* suggests that at every point of the discourse, i.e. in every sentence, each of these objects can be assigned a certain level of *activation*, or *salience*.

One can assume that all the objects referred in a discourse are taken from some *stock of knowledge* shared between the speaker and the hearer (or, in case of a written text, the author and the reader). Then we can regard this set of objects rather as a stack, bearing the most activated items on the top. When an object is mentioned in a sentence, it is moved to the top of the stack (or closely to it, depending on the usage of the referring expression in the sentence). Then, if not referred in the following sentences, it slowly descends, pushed down by the objects which are mentioned in these sentences. Given this model, the quantity of *salience* of an object determines how high this object is located on the stack.

Assumptions have been made (Hajičová, 2003) that if the salience values of the referenced objects in a discourse could be determined, one would be able to induce various characteristics of the discourse. One of them is observing a segmentation of the discourse according to groups of momentarily salient objects along with the identification of their topic(s). Another one could be prediction of a grammatical form of the referring expressions (or, more generally, their strength), eg. pronominal vs. noun referent. Some of these assumptions will be addressed and analyzed in this work.

2.3.1 Salience algorithm

A deterministic procedure to determine the salience values of the coreference chain in the flow of a discourse on a sentence-by-sentence basis was introduced by Hajičová et al. (2006). Its evaluation was presented on one sample document only, because not much data with the necessary annotation were conveniently available at that time. However, the results of the algorithm were also visualized to provide more human-readable feedback.

Let us recall the salience algorithm, as defined by Hajičová et al. (2006) – consider the following situation: An object x represented by the referent r has the salience degree $dg_x^n(r)$ after the n-th sentence of a document is uttered. Then,

the salience value of the object x is defined after its first mentioning by a linear sentence-by-sentence processing as follows:

After each sentence, the salience degree of the object x is modified:

1. $dg_x^n(r) = -1$ if r carries TFA value t or c in the n-th sentence,
2. $dg_x^n(r) = 0$ if r carries TFA value f in the n-th sentence,
3. $dg_x^n(r) = dg_x^{n-1}(r) - 2$ if r is not included in the n-th sentence and has been mentioned in the Focus of the last (not necessary immediately) preceding sentence ($(n-1)$-th through 1st sentence),
4. $dg_x^n(r) = dg_x^{n-1}(r) - 1$ if r is not included in the n-th sentence and has been mentioned in the Topic of the last (not necessary immediately) preceding sentence ($(n-1)$-th through 1st sentence).

Note that this formulation of the salience algorithm does not define the salience value of x before it is first mentioned in the document.

The salience algorithm distinguishes between the Topic/Focus dichotomy and the TFA attribute values ($c/t/f$), according to the theoretical background summarized in 2.2. However, in the scope of this work, we will make a simplification at this point and use the term Focus synonymously to the TFA value f and likewise Topic synonymously to the TFA values c or t. The reasons are rather of technical nature; although a heuristic algorithm proposed by Sgall et al. (1986) has been stated and tested by Hajičová et al. (2005) for "converting" the $c/t/f$ values to Topic/Focus, its results were not fully deterministical. Furthermore, this algorithmic procedure could not be reproduced within the scope of this work.

2.4 Decision Trees and Random Forests

2.4.1 Decision Trees

The decision tree classifier is one of widely used machine learning algorithms. The model of this classifier is a tree graph in which each non-leaf node represents a decision about one feature and the branches leading from this node correspond to decisions about the feature values. Finally, the leaf nodes are labeled by the target feature values. Having this kind of tree, the classification of a new instance is quite simple – its target class is determined by the leaf node of the path beginning in the root and following the decisions on the way down according to the features of this instance.

Learning of this kind of tree parametrization is however not so simple, some questions occur when choosing the attributes which should contribute to the classification and their order. We will demonstrate it on the basic tree learning algorithm called *ID3* (Quinlan, 1986). This basic algorithm is a greedy search top-down algorithm, beginning in the root. When building a new node, the question is: *"Which attribute should be tested in this node?"*, which could be also stated as *"Which attribute classifies the training examples in the best way?"*. This question is answered by a statistical test, in particular (in the case of *ID3*) by the property called *information gain*. This criterion expresses the reduction of entropy caused by partitioning the examples according to the chosen attribute. The entropy of set S relative to the c-class classification is defined as follows:

$$Entropy(S) := \sum_{i=1}^{c} -p_i \log_2 p_i \tag{2.1}$$

where p_i is the proportion of S belonging to class i. In general, the entropy can be seen as some "indefiniteness", or "vagueness". Then the definition of the information gain looks like this:

$$Gain(S, A) := Entropy(S) - \sum_{v \in Values(\mathrm{A})} \frac{|S_v|}{S} Entropy(S_v) \tag{2.2}$$

where S_v is the subset of S for which attribute A has value v. During the tree construction, at each node, this value is computed for each attribute and the attribute with the highest information gain is chosen for the split. However, there are also other measures which can be used as a split criteria instead – for example gain ratio (in principle an adjusted information gain, without the preference of many-valued attributes).

The above presented *ID3* algorithm was later enhanced by Quinlan (1993) under the name *C4.5*, namely by allowing also for continuous variables (using a threshold splitting), handling missing values and the ability of post-pruning the tree.

Another approach to the decision tree learning was introduced by Breiman et al. (1984) and is known as *CART* (Classification And Regression Tree). Coming from a different theoretical background (statistics), the split criterion is stated as the loss of *impurity* created by the split. The usual measure of impurity is *Gini impurity*, but entropy can used as well.

Another important matter in the tree construction is how to avoid the overfitting to the training examples. This is handled by the techniques called prepruning and post-pruning, which are generally methods for regularization of the

model. The idea of the pre-pruning method is stopping the top-down tree building before the overfitting becomes imminent. It could be achieved either by prior setting of the tree depth or by specifying a minimum rate for the splitting criterion – allowing to split only when this threshold can be surpassed. Post-pruning, on the other hand, works with the constructed tree, it simply cuts off parts of it. Usually, the tree is converted into set of (conjunctive) rules first, each corresponding to one path from root to a leaf, the post-pruning can then be done also manually.

2.4.2 Random forests

Random forests are generally a less interpretable, but more efficient extension to the decision tree principle, based on the idea of ensemble learning. Introduced by Breiman (2001), it has recently become very popular model due to its robustness and retaining the expressivity for measuring variable importance. The classifier combines a given number of decision trees, each trained on a subsample of data and on a subsample of features, a combination of bootstrap aggregating (or "bagging") and similar subsampling of features (sometimes called "feature bagging"). It is shown that with a large number of trees, this approach minimizes the noise-sensitivity of the model while simultaneously retaining feature-oriented correlation between the trees, which both contributes to minimize the generalization error. When evaluating new data, the individual decisions of these trees can then be aggregated for regression tasks by averaging their prediction, and for classification tasks by a simple majority voting. Yet another approach for the classification tasks is averaging the probabilistic decisions of the individual trees – this principle is used by the `scikit-learn` library and employed also in our experiments.

Another important feature of the random forests for both the model and problem analysis is the ability to compute importances of the variables used. This calculation is based on summing the split criterion value at each node where the given feature is used, weighted by the number of samples split in this node. These results are averaged over all estimator trees, ensuring robustness of this approach. Therefore, the resulting value summarizes how often and how well the given feature contributed to the splitting among all the trees and their nodes.

3. Data and Tools

3.1 Data Sources

3.1.1 The Prague Dependency Treebank Family

The *Prague Dependency Treebank* (PDT)[1] (Hajič et al., 2006) represents currently the largest annotated corpus of Czech language. The texts are syntactically analyzed using the dependency approach with the main role of the verb. The annotations go from a morphological layer through an intermediate syntactic-analytical layer to a tectogrammatical layer (the layer of an underlying syntactic structure). The process of annotation was performed in the same direction, i.e. from the simplest layer to the most complex. This fact corresponds to the amount of data annotated on each level – 2 million words have been annotated on the lowest morphological layer, 1.5 million words on both the morphological and the syntactic layer, and 0.8 million words on all three layers.

The format of the files containing the annotated data of the PDT family (since PDT 2.0) is called the Prague Markup Language (PML) and is based on XML. Each document data consists of four XML files (typically compressed), one file with the tokenized documents only, each of the rest corresponding to one layer of the annotation and referencing the layer directly superior. Thus e.g. the tectogrammatical layer, as the deepest one, does not contain any surface word forms or purely morphological information itself, but these are accessible through references.

Since PDT 2.0, several updated versions of the treebank were released, specifically PDT 2.5 (Bejček et al., 2011), Prague Discourse Treebank (PDiT) 1.0 (Poláková et al., 2012) and PDT 3.0 (Bejček et al., 2013), whose data is used in this work.

3.1.2 PDT 3.0

The *Prague Dependency Treebank 3.0* (PDT 3.0)[2] is a gradual extension upon the PDT 2.0. It represents a new manually annotated layer of language description, above the existing layers of the PDT (morphology, surface syntax and underlying syntax) and it portrays linguistic phenomena from the perspective of discourse

[1] `http://ufal.mff.cuni.cz/prague-dependency-treebank`
[2] `http://ufal.mff.cuni.cz/pdt3.0`

structure and coherence. Also other various types of information were added and the annotation on all layers was further revised and fixed. However, from the view of this work, the following newly added types of information are important:

- extended textual coreference;
- bridging anaphora;
- pronominal textual coreference of 1st and 2nd person;
- genres of documents.

All the newly added annotation mentioned above was performed on the tectogrammatical trees and technically is a part of the underlying syntax layer (*t-layer*) of the PDT.

With its 49,431 manually annotated sentences from Czech newspapers, the treebank serves as a large-scale resource for linguistic research in the area of discourse analysis as well as for computational experiments concerning automatic text analysis, information extraction, text summarization and other branches of NLP research.

Figure 3.1 (taken from the Prague Discourse Treebank annotation manual) visualizes the tectogrammatical tree structure of one sentence, along with an arrow visualization of the coreference relations. The notation also distinguishes the grammatical and textual reference and includes a bridging anaphora relation. Each tectogrammatical node (or simply *t-node*) has its attributes visualized, such as its tectogrammatical lemma ("potřebovat" – *"to-need"*), functor ("ACT", "PAT", "PRED",...) or a specific sub-type of its reference relation ("SPEC", "WHOLE_PART"). Also note that there are some t-nodes added without any counterpart in the surface representation – such as the root node of the sentence. Another examples would be technical nodes generated e.g. in places of naturally elided expressions, such as zero pronouns. On the other hand, some surface tokens are not represented in the tectogrammatical structure, such as prepositions or auxiliary verbs, their function in the sentence is captured by attributes of the existing t-nodes.

The Prague Dependency Treebank 3.0 is the only source of linguistically annotated data used for the purposes of this work.

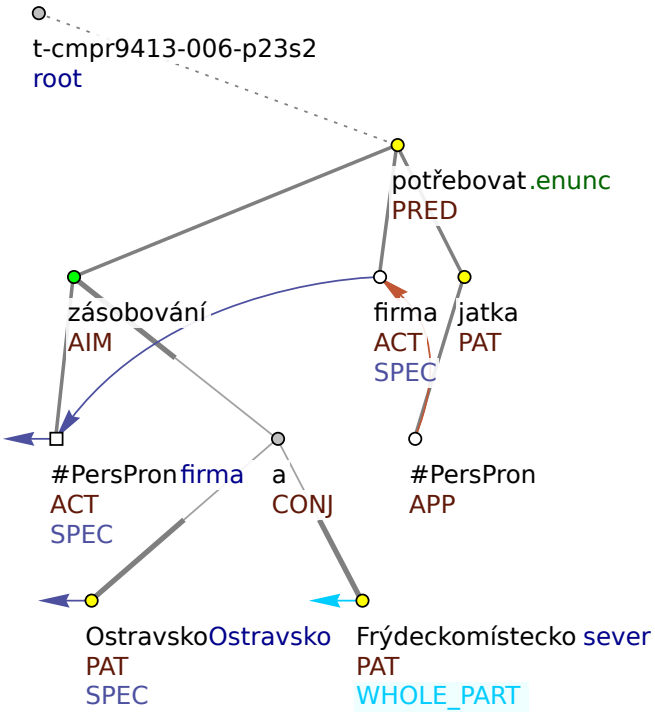


Figure 3.1: Example of coreference annotation for the following sentence: *Pro zásobování Ostravska a Frýdeckomístecka potřebuje firma svá jatka. (The company needs its slaughterhouse in order to supply the Ostrava and Frydek-Mistek regions.)* The brown arrow is used for a grammar coreference relation, navy blue arrows for textual coreference; turquoise arrow for bridging reference (to an expression in another sentence).

3.2 Training and Test Datasets

PDT has already prepared 3 groups of datasets according to the data partitioning typical for the NLP tasks: the training data, the development test data and the evaluation test data. The training datasets cover approximately 80%, development 10% and evaluation 10% of the whole set of data (these proportions hold for all the three layers of annotation).

In this work, we exploit the prepared partitioning of PDT, but we use only the training dataset. Furthermore, for preliminary experiments and some of the more time-consuming tasks, we use only one eighth of the whole training data, the part denoted *train-1*. Throughout this work, we will often refer to this smaller subset as *train-1*, in contrast to the whole training set, denoted as *train-all*.

For a more detailed quantitative analysis of the datasets from the perspective of the features investigated in this work, see Section 4.1.1.

3.3 Tools

3.3.1 Tools for PML

For the batch-processed salience analysis, more convenient data browsing and other manipulation, several tools were used:

- `btred`[3] – Perl-based interface for macro scripting specialized on processing the PML data. Created as a tool for PDT 2.0 (thus applicable also on PDT 3.0), and used in this work as a main instrument for manipulating with the PML data and processing them. The main part of the salience algorithm is implemented by means of `btred`.
- `Tree Editor TrEd`[4] – a viewer and editor of the PDT annotation files, part of the PDT 2.0 distribution. Additional plugins were installed for handling the extra attributes, e.g. coloring of the coreference chains.
- `XSH2`[5] – XML editing shell, used for the extraction of lemmata from the PDT XML format.

[3] `http://ufal.mff.cuni.cz/~pajas/tred/btred.html`
[4] `http://ufal.mff.cuni.cz/~pajas/tred/index.html`
[5] `http://xsh.sourceforge.net/`

3.3.2 Classification and Regression Tools

- `scikit-learn`[6] (Pedregosa et al., 2011) – an open source Python library covering large amount of machine learning tasks and related data processing, data analysis and visualization tools. Used extensively in Chapter 5.

3.3.3 Tools for Clustering

- `pygraphviz`[7] – a Python interface to the Graphviz (Gansner and North, 2000)[8] open source graph layout and visualization package. It was used here for the visualization of document clustering in Chapter 6.

3.3.4 Miscellaneous

- `R`[9] – the R language for statistical computing was used for plotting the salience graphs.
- `Perl`[10] programming language – used for some simpler text and data manipulation.
- `Python`[11] programming language – used for various more complicated data manipulation, as well as plotting some of the bar charts.
- `LibreOffice`[12] `Calc` – a spreadsheet program used for manipulating and plotting especially the data of salience leap heights.
- various Unix shell scripts and makefiles – for smaller tasks, especially for the purposes of batch execution of the repeated ones, some simple scripts were written for the purposes of this work. These tasks included especially the output evaluation, but also grid-searching for parameters or format conversions and adaptations of the data. All these scripts are also present as a part of this work on the enclosed CD-ROM (see Section 6.3).

[6] `http://scikit-learn.org/`
[7] `http://pygraphviz.github.io/`
[8] `http://www.graphviz.org/`
[9] `http://www.r-project.org/`
[10] `http://www.perl.org/`
[11] `http://www.python.org/`
[12] `http://www.libreoffice.org/`

4. Salience Analysis and Interpretation

4.1 Sentences, Coreference and TFA Statistics

Before we proceed to analyze the salience models and its behavior, we should present some statistics about the data and the features which the salience is built upon. Also, the quantitative characteristics of the documents at hand may be useful in the later part for a further analysis during the experiments.

4.1.1 General and Sentence Statistics

Table 4.1 presents an overview of general quantitative characteristics for both training sets used further in the experiments.

	train-1	*train-all*
No. of documents	316	2533
Total no. of sentences	4700	38727
Avg. no. of sentence per doc.	14.9	15.3
Total no. of t-nodes	68626	567258
Avg. no. of t-nodes per sentence	14.6	14.6
Avg. no. of t-nodes per doc.	217.2	223.9

Table 4.1: General statistics of the datasets.

More detailed distribution of the counts of sentences per document is shown in Figures 4.1 and 4.2. Note that the most typical per-document sentence count in both cases is 8, which is far below the average value.

4.1.2 Coreference

Perhaps the more important one of the two main pillars which the salience concept is built upon, is the concept of the coreference relation. To understand the salience models, we have to explore first the basic characteristics of the coreference chains themselves in our data.

The counts of grammatical and textual coreference links in *train-1* and *train-all* are summarized in Table 4.2 and Figure 4.3 along with the counts of bridging anaphora links. Those are not coreference relations in the strict sense, but since

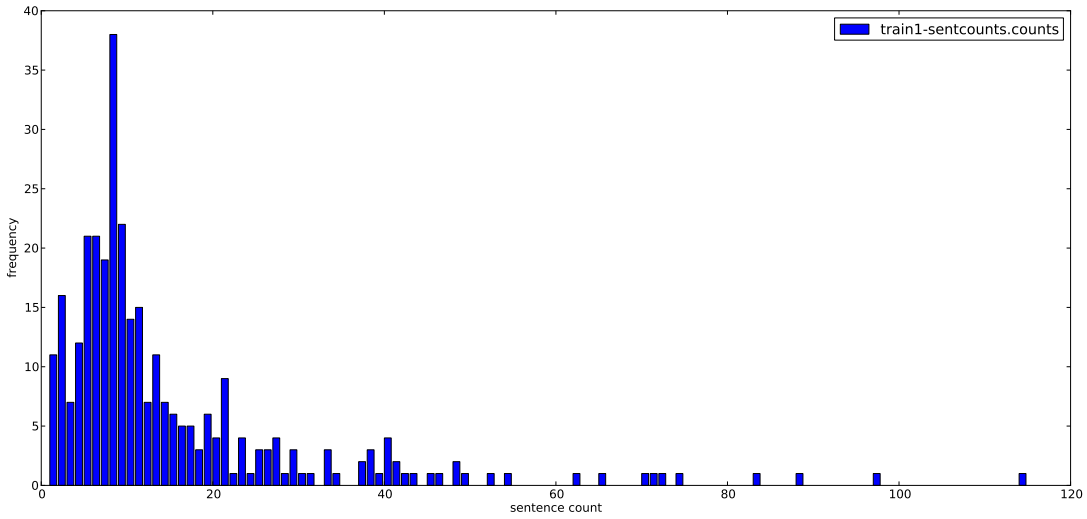


Figure 4.1: Distribution of the per-document sentence counts in *train-1* dataset.

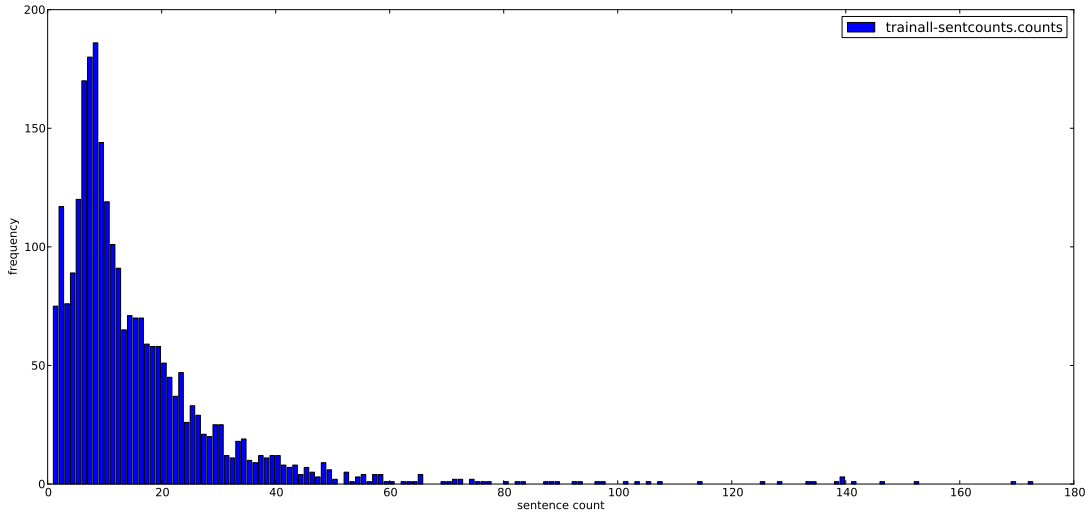


Figure 4.2: Distribution of the per-document sentence counts in *train-all* dataset.

coreference type	*train-1*	*train-all*
grammatical	2226	18156
textual	7514	67535
bridging anaphora	1987	23512

Table 4.2: Coreference type link counts

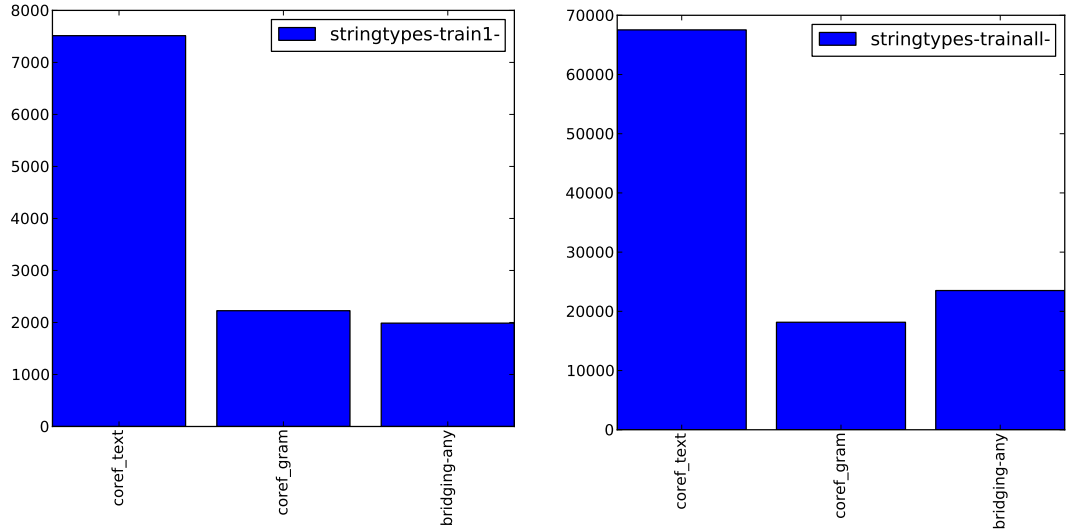


Figure 4.3: Counts of the coreference link types in *train-1* and *train-all* dataset.

we have experimented with using them as such (see Section 4.1.2), the numbers are listed there for comparison.

Another important data are the counts of the whole coreference chains, presented in Table 4.3. These frequencies, especially when related to the number of documents, might be interesting in some analyses concerning the comparison of the word counts with the coreference chains in Chapter 6).

	train-1	*train-all*
No. of documents	316	2533
Total no. of coref. chains	4519	39415
Avg. no. of coref. chains per doc.	14.3	15.8

Table 4.3: Counts of the whole coreference chains in the datasets, related to numbers of documents.

Chain lengths When speaking of the length of a coreference chain, we have adopted the definition of *coreference chain length* being the number of coreference nodes (i.e. co-referring expressions) in the chain. Thus the most frequently appearing chain has length of 2, meaning two anaphoric expressions referring to

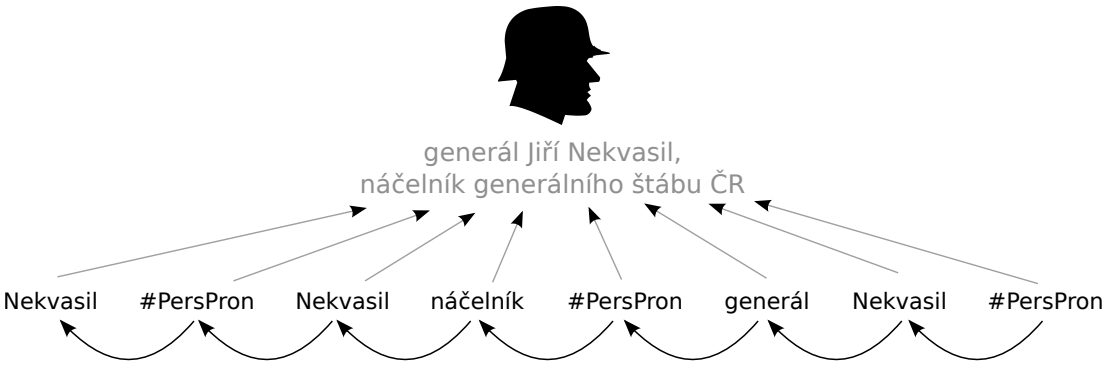


Figure 4.4: Example of a chain of 8 t-nodes co-referring to one real-world entity (*general Jiří Nekvasil, Chief of the General Staff of the Czech army*). Thus, the length of this chain is 8, according to our definition.

the same item (in the PML representation, this is represented by two tectogrammatical nodes with one coreference relation between them, typically the first one being the antecedent of the second one). An example of a coreference chain of length 8 is shown in Figure 4.4. According to our definition, we have acquired the length-frequency figures presented in Figure 4.5. The distribution is not suprising; the chain length of 2 coreferents is the most typical case, whereas the frequency of longer chains drops rapidly. However, although the "tail" of the graph was cut off for sake of readability, the longest chain encountered in the data was 89 nodes long (and it was found in a document of 114 sentences). To complete the data, we will add that the average length of a corefence chain in *train-1* is 5.1.

Adding Bridging Anaphora The coreference chains are the main platform for the salience analysis and modeling of a text. If the salience should be used to model the dynamics of e.g. some inherent topics of the text, it would be convenient to have at our disposal the coreference chains "as long as possible". In other words, one should make effort to identify as many connecting relations between associated expressions as possible. In this pursuit, we have experimented also with using the annotation of bridging anaphora as an additional source of coreference relations. The experimental approach was quite straightforward; since the salience algorithm does not distinguish between types of coreference, we can let it treat the bridging relations the exact same way as the "regular" coreference.

However, when commiting to this step, one has to bear in mind that the bridging relations does not have so "strict" characteristics, which can, to a certain degree, also affect the results of the subsequent salience modeling. For instance, if one coreference chain contained more than one *set-subset* or *contrast* bridging relations, there is no guarantee that the referenced item would stay the

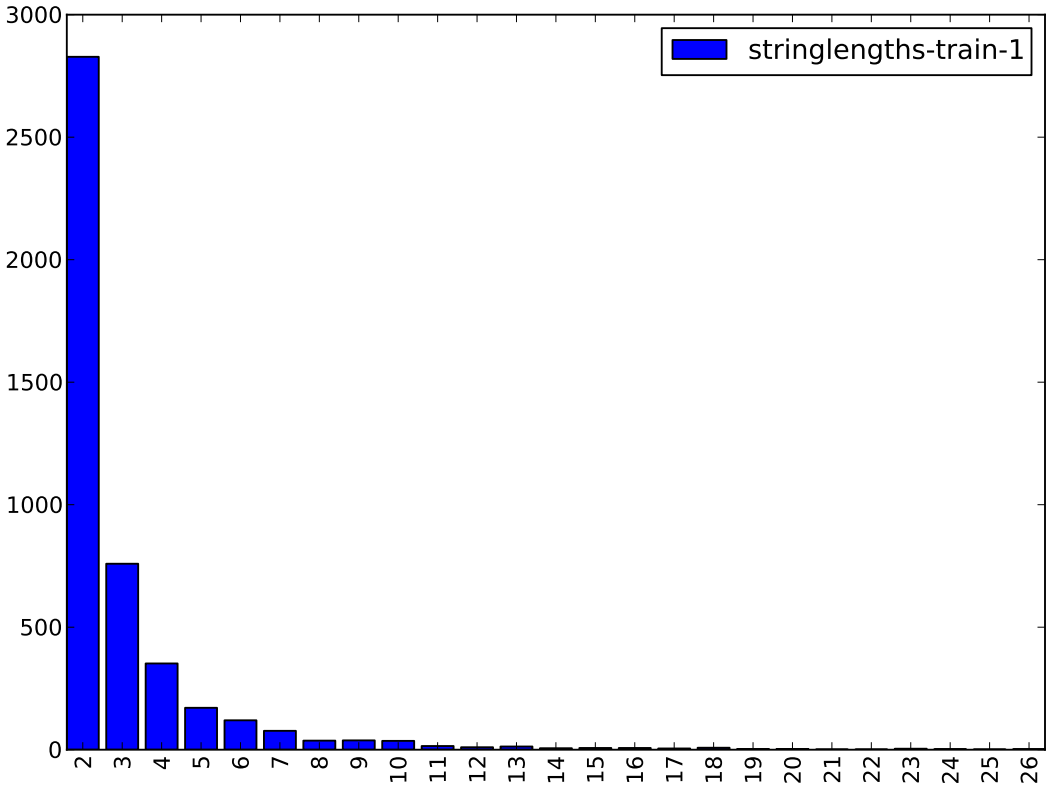


Figure 4.5: Frequency of lengths of coreference chains in *train-1* dataset; cut off at length of 26 nodes.

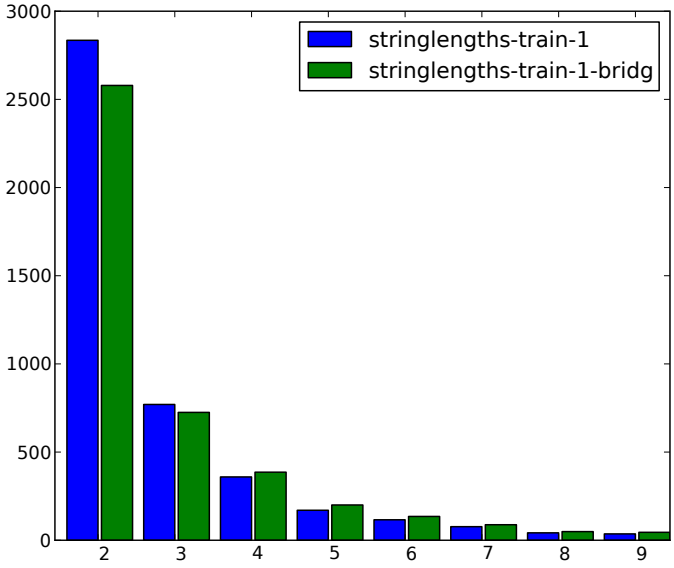


Figure 4.6: Frequency of lengths of coreference chains in *train-1* dataset – the impact of adding bridging anaphora.

same throughout such semantic shifts. The measure of this effect can be hardly anticipated – ideally, one would have to perform two sets of all the planned experiments and maintain two sets of results, comparing them and evaluating the differences continuously.

Furthermore, when we examine the actual impact on the length of the coreference chains (see Figure 4.6), the influence is obvious, but not as large as we presumed. Taken into account the above objections, we have finally decided to abandon this path in the scope of this work and perhaps leave it for a further investigation.

4.1.3 TFA

The proportion of the TFA markers for the tectogrammatical nodes in *train-1* dataset is visualized in Figure 4.7. In accordance to the PML annotation customs, '`t`' stands for a non-contrastively contextually bound expression (represented by the node), '`c`' for a contrastive contextually bound expression and '`f`' for a contextually non-bound expression. Finally, the '`-`' bar in the chart represents the amount of nodes not marked with any TFA value.[1]

[1]These are mostly technical cases, e.g. root of the tectogrammatical tree or of a paratactic construction, or a foreign-language expression, which has often a special treatment in the PML annotation scheme.

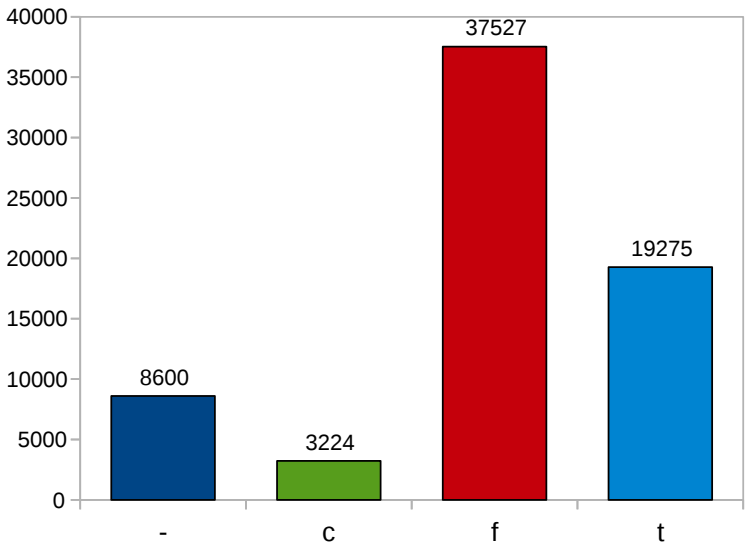


Figure 4.7: Frequency of TFA values in *train-1* dataset.

4.2 Salience Graphs and Interpretation

4.2.1 Salience Graphs

Figure 4.8 presents an example of a salience graph for a short document. The graph was generated from the Czech original of the document, the presented English translation tries to preserve partially the original sentence structure. In the chart, each coreference chain is represented by a numbered polyline, the members of the chain are marked by the corresponding color in the text.

4.2.2 Salience Graph Generation Procedure

One of the main parts of this work was to automatize the procedures needed for the visualization of the salience for each document. This consists of several steps, the whole process being summarized in Figure 4.9. Each step is performed by a procedure in a script file, making also the intermediate results analyzable. Each of these script files uses a programming or scripting language which seemed the most appropriate for the task: When working with the PML files, `btred` is used, `Perl` itself is employed for non-PML text manipulations, and `R` language was chosen for the graph visualization part. Some parts of the scripts were originally created during the preparation of Hajičová et al. (2006). However, their ad-hoc nature made them largely impossible to suit our purposes, thus they were all significantly rewritten, made more readable, documented, and hopefully reusable.

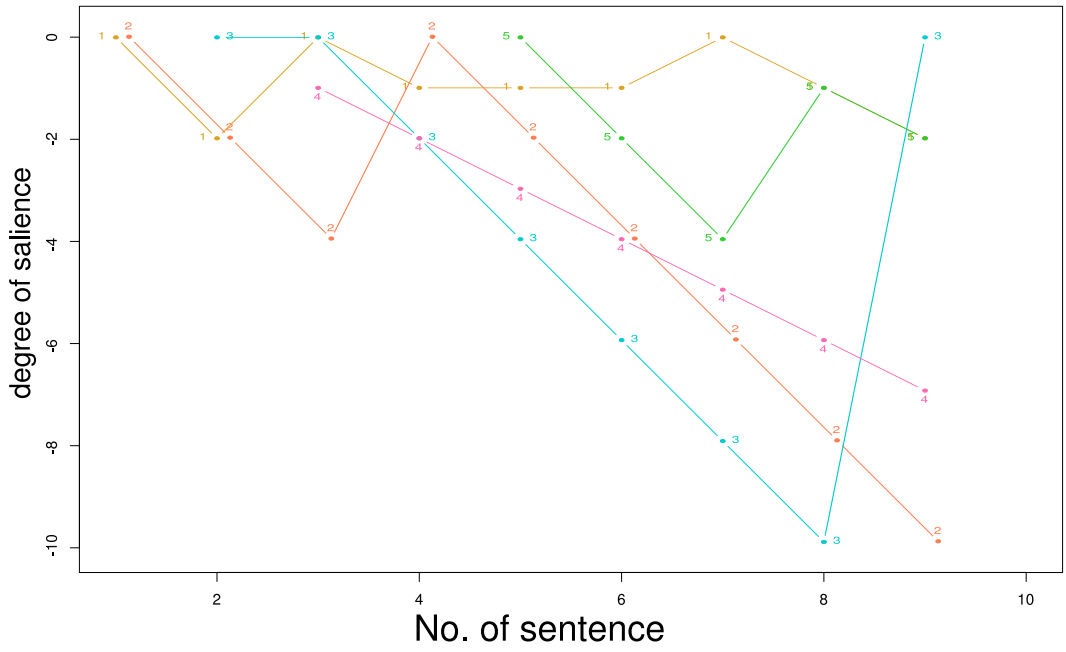


Figure 4.8: Example of a short document from PDT along with its salience graph.

(1) *Accounter and one million have disappeared*

(2) *Brno*

(3) *Since 11th June, when (he) left the work around 3 AM and did not come home, the police is searching for a 27-year-old Stefan Misik, main accounter of casino 777 on the Svobody square in Brno.*

(4) *The searched-for man had over million crowns with him and could be a victim of a violent crime.*

(5) *Stefan Misik resides in Pradlacka street and has a well-built, 178-cm-high figure, short brown hair and a pea-sized birthmark on left side of his neck.*

(6) *During the speech, (he) burrs.*

(7) *Last time (he) was wearing (on him) a bright shirt, black jeans and brown loafers.*

(8) *On the neck, (he) was wearing a silver chainlet with a sign of Cancer, in a black bag had also a new passport and magnetophone tapes.*

(9) *Witnesses can report to the nearest police office, the 158 (phone) line or the I. department of Crime Service in Brno, phone 05/4116 2525.*

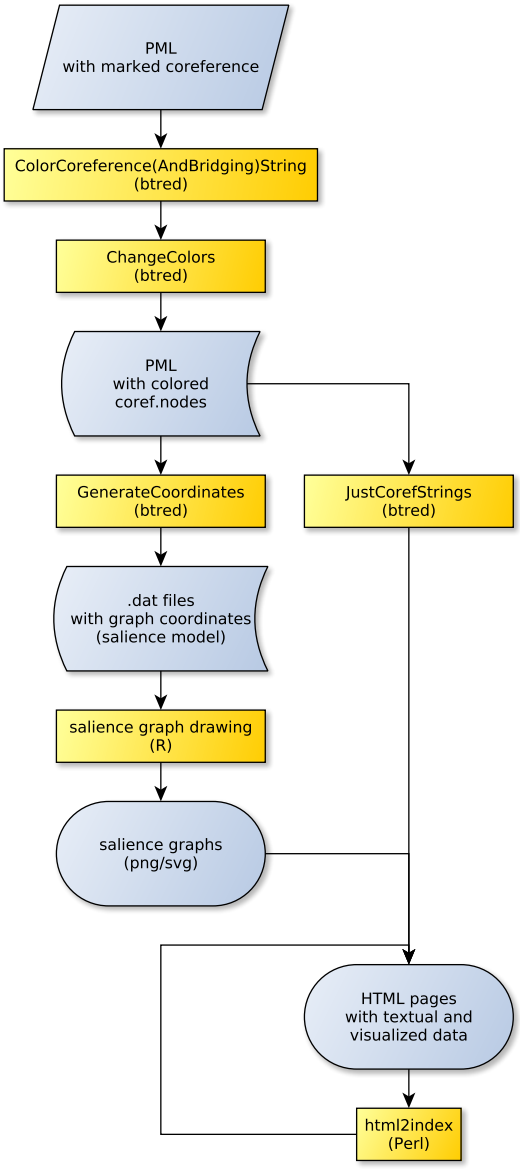


Figure 4.9: Flowchart of the data processing from PML corpus data to the salience graph visualization embedded in an HTML page.

The first step is to modify the PML files by identifying the nodes of each coreference chain and marking them accordingly – we call this process "coloring the coreference chains". This is achieved by applying a simple algorithm of linearly going through the tectogrammatic tree nodes, inspecting their direct coreference antecedents and denoting them by the corresponding color number identifier. The next small step, rather technical, is to order the color identifiers sequentially with respect to the linear flow of the sentences (this process actually is not necessary for the functionality, rather a convenience for further inspection).

Computing the salience degree of each coreference chain members is done in the subsequent step. This is where the salience algorithm is applied to the colored nodes. In each sentence in the "colored" PML files, salience degree is computed for each coreference chain which has appeared so far, and extracted into an external file. This information, serving later as "coordinates" in the final graph, is then fed into R script described further.

The actual graphical form of the salience graphs is generated by a script in the `R` programming language. As its input, it is given a set of files (each corresponding to one document) with salience "coordinates": for each occurence of a coreference chain member, there is a line in the file with its coreference chain identifier, sentence number and the salience degree of the member's occurence. From this coordinates of all points of the salience graph, a graphical file is generated – either as a bitmap (PNG) or in a vector-based format (SVG or PostScript). The output is made as readable as possible, providing both colors and numbers for each coreference chain curve, as well as a slight shifting of the curves to reduce their overlaps. However, the variability of the salience behavior of the chains, inherent density of the curves in a large part of the documents and the variability of the documents' lengths make it hard to effectively generalize some of the techniques used for improving the readability and clarity of the graphs.

4.2.3 Vertical Cut

Moving on the "horizontal" axis on the graph, i.e. sentence by sentence, and observing the current trend of all the chains at once, certain vertical breaks can be identified in the salience models. These suggest a slight change of topic in the particular sentence, where several new objects emerge or re-activate and the old ones fade away. From this point of view, the salience models could be used e.g. for automatic segmentation of previously unsegmented text by "cutting" the text at this breaks, perhaps into paragraphs. Furthermore, the objects emerging at the identified breaks (or later in the beginning segment) can suggest the topic of the

current segment. The design of an actual algorithm for such automatic process is not covered by this work, although one should be able to test such algorithm rather conveniently on the PDT data with the original paragraph segmentation preserved, thus applicable as the gold standard.

4.2.4 Horizontal Cut and Leap Height

Another approach to the models would be to draw one or more horizontal lines in the graph to mark a certain level of salience. One can assume that these levels can express the amount of activation an object must have to be referred to by certain grammatical means – a weak or zero pronoun is expected to refer to an object with high activation, whereas less salient objects are re-activated by more specific expressions, e.g. a definite noun phrase.

To verify these hypotheses, let us introduce a new quantity: *salience leap height*, or simply *leap height*. Each time an object (represented by its coreference chain, i.e. chain of expressions referring to it) is mentioned in a sentence, the *leap height* value indicates the difference of its current salience level and its salience level in the previous sentence. More rigorously, let the leap height value of an object x (or, from another point of view, of its coreferents' chain) in sentence number n (where x is mentioned) be defined as such:

$$LeapHeight(x, n) := dg_x^n - dg_x^{n-1} \tag{4.1}$$

Note that this definition contains not only the "depth" from which the mentioned object emerges, but also takes into account the TFA value of the current referring expression, in the form of its current salience value – being it either 0 or -1. This reflects the idea of differentiating the referent's actual sentence function. This differentiating is proportionally more important with the smaller leap heights and losing its importance with their higher values, which may not necessarily be harmful. This property also results in a possibility of leap height having a zero value, or even a negative value, specifically -1; when the last reference of x was in the focus (had TFA value f) of the previous sentence and the current reference is in the topic (has TFA value t or c). This situation is actually quite common in the discourse; it corresponds to the usual case of a newly emerged object in the $(n-1)$-th sentence, which is subsequenly referred to in the n-th sentence, serving there as a "starting point" (a topic, in the TFA terms). The Example (1) (from PDT 3.0) illustrates exactly this situation – the entity "Ministerstvo hospodářství" (*"Ministry of Economy"*) is referenced once in each of two subsequent sentences (by the same surface form in this case). The first reference is contextually unbound

(marked by the TFA attribute value f), thus bearing salience value 0 (indicated by the subscript number). Subsequently, its occurence in the second sentence is non-contrastive and contextually bound (TFA value t), gaining salience value -1, which results in $LeapHeight(\text{“Ministerstvo hospodářství”}, 2) = -1 - 0 = -1$.

(1) 1. Zkušenosti *Ministerstva/f/*$_0$ hospodářství ČR z loňského roku ukazují, že vzhledem k postupnému zlepšování informovanosti podnikatelů o programech podpory se podstatně zvýšil i jejich zájem o získání finančních dotací od státu.
2. Výsledkem byl značný převis poptávky nad celkovými možnostmi, tedy prostředky, které *Ministerstvo/t/*$_{-1}$ hospodářství dávalo k dispozici podnikatelům prostřednictvím Českomoravské záruční a rozvojové banky.

(*1. The experience of* Ministry/f/$_0$ *of Economy of the Czech Republic from last year shows that due to a gradual improvement of awareness of businessmen about support programs, their interest in public financial grants has grown substantially as well.*
2. The result was a considerable excess of demand beyond the capabilities, or resources, which the Ministry/t/$_{-1}$ *of Economy made available to businessmen via the Českomoravská Guarantee and Development Bank.*)

All the leap-height charts presented in this section has their values normalized to sum up to 1 within the given feature value (TFA or *sempos*). The reason is that in these analyses, we are mostly interested on the distribution within the given feature value, rather than directly comparing the two absolute values at any fixed leap height.

Leap Heights and TFA Figure 4.10 shows the frequency of the leap heights depending on the TFA value of the referring expression. A general rule may be stated that shorter leaps are typical for mentioning in topic (c/t), while the longer ones are slightly more common for mentioning in focus (f).

Also note the fact that the leaps to the topic are apparently more frequent for the odd leap heights, whereas the focus “destination” favors the even leap heights. This is an inherent property stemming from the inclusion of the TFA in the definition of the leap height.

Pronominal vs. denominating referents Let us return to the above mentioned hypothesis about the grammatical form of referents typical for certain salience ranges. Thanks to an elaborate system of the tectogrammatical layer

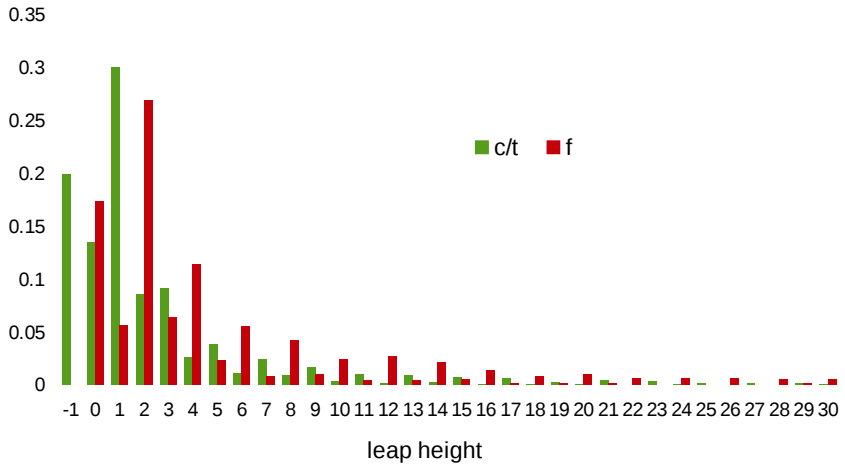


Figure 4.10: Proportions of leap heights comparing the coreferents' TFA values; from *train-1* data. (The y-axis units are ratios of leap heights for the given TFA value normalized to sum up to 1.)

annotation in PDT, we can use the t-node attribute *sempos*[2]. The pronominal expressions are marked with *sempos* value containing `.pron.` (e.g. `n.pron.indef` standing for "indefinite pronominal semantic noun"), whereas the *sempos* value of the denominating expressions contains `.denot.` (e.g. `n.denot` means "denominating semantic noun"); the rest being only quantificational expressions and verbs. With this division, we can visualize the proportions of the leap heights within each of these *sempos* categories in Figure 4.11.[3]

From the chart, it is obvious that there is some disproportion in the behavior of the pronominal referents in comparison to the denominating ones. The quick drop of the pronominals' values beyond the leap height of 1, along with the rather steady decline of the denominators, seems to confirm the declared hypothesis. However, the dominance of the -1 value is quite surprising and calls for a deeper analysis.

The Figure 4.12 thus focuses only on comparing demonstrative and personal pronouns (*sempos* values `n.pron.def.demon` and `n.pron.def.pers`, respectively), because these two are by far the most frequent types among the pronominal coreferents. The difference between them is apparent: while the demonstrative pronouns almost fails to refer beyond the leap height of 1 and serves mostly for the -1-leap reference, the personal pronouns, although also "specialized" on the

[2]From the PDT t-layer annotation manual: "The sempos attribute (semantic part of speech) contains the information regarding the membership of a complex node in a semantic part of speech." (Hajič et al., 2006)

[3]Although the leap height values goes as far as 172, the tail is long and its values neglectable for our purposes – thus the charts are often cut off at the leap height value of 30.

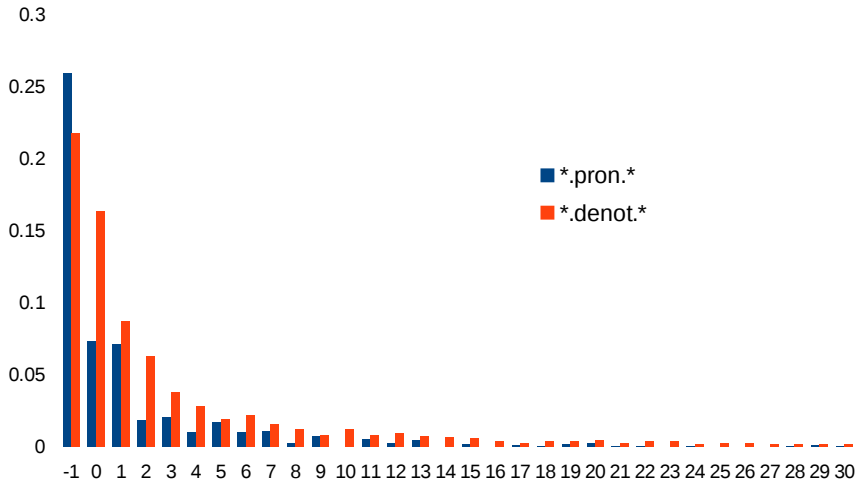


Figure 4.11: Proportions of leap heights for the chosen *sempos* categories; from *train-1* data. (The y-axis units are ratios of leap heights for the given category normalized to sum to 1.)

low leaps, perform best for the leaps of 1 or 0. From this comparison, it is also evident that the demonstrative pronouns were almost fully responsible for the high values of leap height -1 for pronominals in the previous categorial comparison.

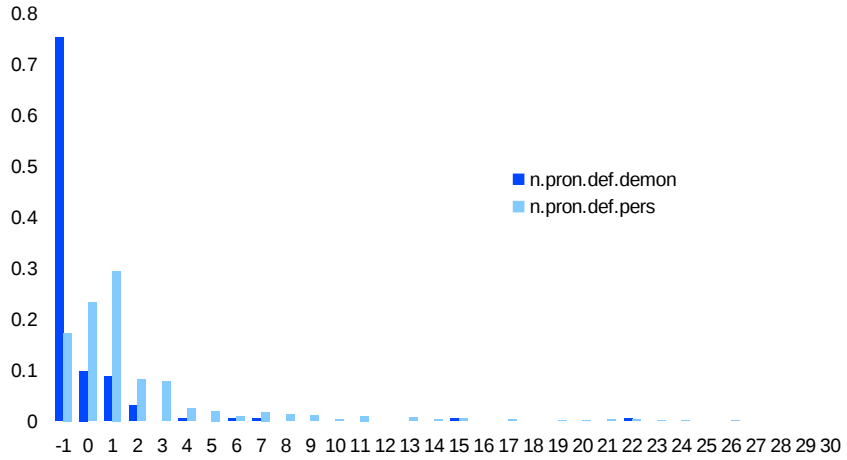


Figure 4.12: Proportions of leap heights for the two chosen pronominal *sempos* values; from *train-1* data. (The y-axis units are ratios of leap heights for the given *sempos* value normalized to sum to 1.)

5. Learning Salience

5.1 Motivation

One of the goals of this work is to explore possibilities of employing salience as a feature in a machine learning task. This target usually requires large amount of data, for which manual annotation may not be available. However, the algorithm-based approach to compute salience degrees requires features such as coreference links, whose automatic resolution reliability for Czech data is currently relatively low (Nedoluzhko et al., 2013). Therefore, we have carried out experiments to find out whether the salience degrees (in some form) can be machine-learnable based solely on morphological and syntactic-layer data. Automatic analyses of these two layers are known to have quite high accuracy on Czech, with morphological tagging around 96% (Spoustová, 2008) and best dependency parsers scoring around 86% (Holan and Žabokrtský, 2006; Koo et al., 2010).[1]

Furthermore, some parts of these experiments may be beneficial also in the direction of understanding the notion of salience and forms of its surface expression. Employing a transparent machine learning model, such as decision tree, we might be able to read possible direct linkage between surface forms of the coreferent (or shape of its context) and its salience development, from the features' positions in the tree.

5.2 Experiment Settings

5.2.1 Genre Filtering

Our source of annotated data, the Prague Dependency Treebank, comprises of various (sub)genres of the newspaper articles. Some of them might not be very reliable in terms of text coherence and discourse flow, thus possibly degrading results of our data-based machine learning approach. Fortunately, there is a newly added document-level information about its genre in PDT version 3.0, so we are able to choose only a supposedly reliable subset of the genres and filter the documents accordingly, if found necessary. The Table 5.1 shows what genre subset was chosen for this purposes, along with the document counts of each genre – which was also one of the determining features for their selection (so

[1] For up-to-date results of dependency parsing systems for Czech refer to `http://ufal.mff.cuni.cz/czech-parsing`.

that a reasonable amount of data would remain even after the filtering). Another features indicating the discourse coherence of the genre (such as frequency of discourse relations) were considered, based on a thorough analysis by Poláková et al. (2014).

genre	document ratio	volume ratio	used?
news	40%	28%	+
sport	11%	10%	+
description	8%	12%	+
caption	7%	1%	−
essay	6%	14%	+
collection	5%	4%	−
comment	5%	6%	+
review	4%	5%	+
invitation	3%	2%	−
other	3%	2%	−
topic_interv	2%	5%	−
advice	2%	3%	−
overview	1%	1%	−
survey	1%	1%	−
letter	1%	1%	−
person_interv	1%	3%	−
program	1%	1%	−
plot	0%	0%	−
weather	0%	0%	−
metatext	0%	0%	−

Table 5.1: Overview of genres with their cardinality, marked by usage in our genre-based filtering.

5.2.2 Detecting Possible Referents

If we are trying to predict salience of the items referred to in the text without the coreference links resolved, all we can do is work with any potential referring expression of an item. Thus first, we have to find a way to estimate whether a word is likely to be a (co)referring expression. The most straightforward approach is to base this decision on part-of-speech (POS) of the given word, since this information is found on the morphological layer and is usually very reliable.

Analysis Table 5.2 shows the relationship between POS of a word and his membership in a coreference chain (coloring). The table is divided into two

parts, the left part captures the word counts among documents of all genres, the right one concerns only the documents after the genre-based filtering described in Section 5.2.1.

	all genres			genre-filtered		
	colored	non-colored	% colored	colored	non-colored	% colored
noun	9209	15433	37.4%	7136	11319	38.7%
pronoun	2682	1146	70.1%	1880	815	69.8%
unknown	2792	8944	23.8%	1879	6208	23.2%
adjective	683	8801	7.2%	528	6558	7.5%
verb	345	7921	4.2%	250	5801	4.1%
adverb	178	3701	4.6%	120	2785	4.1%
numeral	79	2350	3.3%	48	1594	2.9%
other	319	4046	7.3%	234	2981	7.3%

Table 5.2: Frequency overview of relations between morphological parts-of-speech and coreference coloring (i.e. chain membership), data from *train-1*.

There is, however, one problem when studying these relationships: the coreference chain membership in the data is defined for the t-layer nodes, but these do not necessarily have an m- or a-layer counterpart. These cases are denoted as *unknown* in the table. But most often, they correspond to an added technical node on the t-layer to fill a valency spot of a governing node – and if they are involved in a coreference relation, it is usually a short-range grammatical coreference link. We cannot capture these nodes in any way in our approach, but for these reasons, we do not lose much.

Based on the given analysis, we will consider the first two POS categories, nouns and pronouns, as the most likely indicators of the coreference membership, i.e. as the probable identifier of a referring expression.

It should also be noted from the given numbers that the genre-based filtering did not change these distributions very much. However, we will stay with the advice from Poláková et al. (2014) to leave out the problematic genres and keep only the more reliable ones. Therefore, all the following experiments (if not stated otherwise) uses the genre-based filter for the data.

5.2.3 Leap Height

Salience development is a characteristic of the whole coreference chain (or, more precisely, its referent). However, while not able to resolve the coreference links, we have to work with the individual referring expressions. In this case, salience

degree at the current word reflects only its TFA value, deeming it to values 0 or −1 only. To capture also more about the previous salience development, the *leap height*[2] variable is more suitable. Thus in the experiments, we will predict the value of leap height instead of the actual salience degree.

Despite this necessary limitation, the predicted value can be still useful in a further application, such as text segmentation (see also Section 4.2.3). In such application, the most important information in this direction would probably be the number of referring expressions in the current sentence and particularly the "strength" of their reference, in the same sense which is stated by the leap height value.

There is one more restriction related to the leap height target value. It is not defined for the first member of each coreference chain and there is no straightforward or clear way of determining a default value for such cases. Probably the best way to solve this problem will be to train the models *without* these first members (they comprise ca. one third of the data, this ratio holds roughly also after the filtering phases described further) and, if these models prove successful enough, use them to predict this value for the left-out nodes. The outcome of such method would be also interesting from the view of the salience concept itself, as this default value is expected to reflect an implicit position of the referred object in the stock of shared knowledge at the moment of its first reference.

Logarithms and bins Leap height is a discrete feature taking integer values from −1 to potentially $2n$, where n is number of sentences in a document. However, in the context of its relationship to the salience degrees, a similar assumption can be made such that the higher the values are, the smaller is the distinction between them. In other words, the difference between leap heights −1 and 2 is quite significant, but the difference between 23 and 26 is actually not very important. From this assumption we can induce two different approaches to transforming the leap height value as a target feature. The first one is to compute a logarithm of the leap height value and apply a regression approach to the problem. The second one would be a classification approach which conveniently puts various leap height values in a specified number of bins (but again of logarithmic size) and takes each bin as one class of the target feature. Both of these approaches were tested in our experiments, with their results presented along each other.

[2]For its definition and introduction, see Section 4.2.4.

5.3 Machine Learning Model

5.3.1 Features

For the task of predicting the leap height value, values of the following set of features were collected from the data:

Surface layer features

- position of the sentence within document – from its beginning
- position of the sentence within document – from its end
- position of the word within sentence – from its beginning
- position of the word within sentence – from its end

When computing the position of the word within sentence, only non-punctuation tokens were considered as words.

Morphological layer features

- word-distance[3] from the nearest preceding noun/pronoun
- word-distance to the nearest following noun/pronoun
- POS+subPOS tag of the one before preceding word
- POS+subPOS tag of the preceding word
- POS+subPOS tag of the current word
- POS+subPOS tag of the following word
- POS+subPOS tag of the one after following word

Analytical layer features

- analytic functor
- analytic functor of the parent node
- analytic functor of the first child node (if present)
- POS+subPOS tag of the parent node
- POS+subPOS tag of the first child node (if present)
- relative position of the word to the predicate[4]

All the features presented so far are commonly used in various machine learning tasks to describe the close context of the word or its position within sentence or document. However, we need also some means to try to locate the possible

[3]This and all the following word-distances may span several sentences. Also, similarly to thw word-position features, only non-punctuation tokens were considered as words.

[4]Word distance on the surface layer; negative value means that the predicate is on the right of the current word and vice versa.

coreference antecedent, as this information is crucial to the salience development and thus consequently to the leap height value. This is a field of *coreference resolution* (CR), where we can exploit some previous research and gather inspiration. Our set of features specialized on coreference were inspired particularly by some ideas from Novák and Žabokrtský (2011) and Soon et al. (2001).

CR-like features

- word-distance from the last word with the same lemma
- sentence-distance from the last word with the same lemma
- word-distance from the last pronoun with number and gender agreement
- sentence-distance from the last pronoun with number and gender agreement
- word-distance from the last noun with number and gender agreement
- sentence-distance from the last noun with number and gender agreement

Feature extraction All the features for the machine learning task were extracted from the PDT data. For the extraction, a `btred` script was used, going through each coreference-colored tree node on the t-layer of the data and trying to find the corresponding m/a-layer node. If there was any and if its POS value corresponded to a noun or pronoun, the rest of its feature values were gathered and this node was added as one data instance. Thus, each data instance corresponds to one noun or pronoun m/a-layer node, whose corresponding t-layer node is coreference-colored. Furthermore, the first node of each coreference chain was filtered out, because the value of leap height value is not defined for such nodes (as described before).

Binarizing categorical features For technical reasons, categorical features (such as morphological POS or analytic functors) are converted each into a set of binary features encoding its value. For this task, the usual *one-of-K* or *one-hot* encoding principle is applied, where a vector of N new binary features is created for N most frequent values of the given original feature, assigned 1 where the instance has this particular value, otherwise 0. Therefore, for each given categorical feature, there is always at most one 1 among this vector of new features, the rest are zeros. An additional position is added to this vector indicating that the original value is other than one of the N listed.

Final feature and instance count After this preprocessing step, 104 features + one target variable are available for training a model. From the datasets

train-1 and *train-all*, 5024 and 46955 instances, respectively, are available after the filtering described before.

Target value transformation The leap height as the target value was transformed in several different ways (according to the ideas in Paragraph 5.2.3), plus the original value was also used in one particular set of experiments. For the classification task, three binning schemes were used:

- *none* (keeping the original leap height value),
- dividing into 7 bins: $[(-1), (0), (1), (2), (3, 4), (5 - 10), (> 10)]$,
- dividing into 3 bins: $[(-1, 0, 1), (2, 3, 4), (> 4)]$.

The choice of these arbitrary schemes was determined by two main reasons: Firstly, to reflect the "logarithmic" nature of the leap height value (or its relative importance), and secondly, to keep the frequency distribution among the bins relatively similar. For the regression task, apart from the analogous *none*, natural logarithm was used – however, the leap height value had to be pre-adjusted by adding 2 to ensure the logarithm to be defined even for the -1 and zero leap height value.

5.3.2 Decision Trees and Random Forests

For the task of machine learning and prediction of the leap height value, two ML methods were chosen: decision trees and random forests.[5] The reason for this choice was mainly the well-known interpretable nature of these model types. Especially decision trees are often used as a white-box model, their structure being analyzed after fitting. On the other hand, the random forests can be also relatively efficient predictors, with their accuracy comparable to other methods in many classification tasks. And at the same time, they can offer estimates of relative feature importance, which is an important property also exploited in this work.

5.3.3 Performance measures

Each experiment—combination of the original dataset (*train-1* or *train-all*), target value transformation and type of model—was evaluated by the process of 5-fold cross-validation on the whole dataset. The average of all 5 scores is presented

[5]For the theoretical background of these methods, see Section 2.4.

along with a ± value indicating the confidence interval boundaries (assuming the error follows the normal distribution).

Classification The main metric used for all the classification tasks was accuracy, i.e. ratio of correct predictions among the testing dataset.

Classification baseline Baseline computation for the classification task was done using the usual way of assigning every instance the target value which was the most frequent among the training examples.

Regression For the regression tasks, two evaluation metrics are presented. The first one is perhaps the most commonly used *Root-Mean-Square Error* (*RMSE*, also known as *root-mean-square deviation*), whose values are however dependent on the units of the actual regression task. If $\hat{\boldsymbol{y}}$ is a vector of n predicted values, and $\boldsymbol{y}$ is the vector of the true values, then the *RMSE* of the predictor is computed as follows:

$$RMSE := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{5.1}$$

It should also be noted that for since the *RMSE* metric is actually an error metric, the lower results are better.

The second regression metric one is the coefficient of determination (R^2), which summarizes the explanatory power of the model more generally; its values are independent on the actual task evaluated. If $\bar{y}$ denotes the mean of true values $\bar{y} = \frac{1}{n}\sum \boldsymbol{y}$, then R^2 value is computed using fraction of sum of square of residuals (regression error, SS_{error}) and total sum of squares of true values (SS_{total}):

$$R^2 := 1 - \frac{SS_{\text{error}}}{SS_{\text{total}}} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.2}$$

From this definition it follows that the higher the R^2 measure is, the better is the result, with the maximum of 1 (in the case when $\hat{\boldsymbol{y}} = \boldsymbol{y}$, i.e. the predicted values are equal to the real ones).

Regression baseline Analogously to the classification task, the baseline for regression was set by predicting always the mean value computed from the training instances. From the definition of such regressor, it follows that since $\hat{y}_i = \bar{y}$, then $SS_{\text{error}} = SS_{\text{total}}$ and thus the R^2 value will be zero for this baseline regressor.

5.4 Results and Discussion

5.4.1 Decision Tree Classification

Classification results of the decision tree models are listed in Tables 5.3 and 5.4, divided according to the dataset used (*train-1* and *train-all*). Two key parameters of the model were determined by a grid-search based on their cross-validation results. Of two possibilities of the splitting criterion, Gini impurity and information gain, the former was consistently better in all experiment settings and thus used in every case. However, the second parameter, maximum depth of the tree, was adjusted for each setting – and is noted in the last column of the tables.

binning	baseline	cross-validation	max_depth
none	0.333	0.449 ± 0.036	5
$(-1), (0), (1), (2), (3, 4), (5–10), (>10)$	0.333	0.509 ± 0.041	6
$(-1, 0, 1), (2, 3, 4), (>4)$	0.693	0.792 ± 0.049	5

Table 5.3: Results of decision tree classifier on *train-1*.

binning	baseline	cross-validation	max_depth
none	0.315	0.465 ± 0.011	8
$(-1), (0), (1), (2), (3, 4), (5–10), (>10)$	0.315	0.535 ± 0.020	8
$(-1, 0, 1), (2, 3, 4), (>4)$	0.657	0.783 ± 0.014	6

Table 5.4: Results of decision tree classifier on *train-all*.

5.4.2 Random Forest Classification

Tables 5.5 and 5.6 show the results of the random forest models on the same classification tasks. The grid-search for the two key parameters was performed similarly to the decision tree classification, obtaining the analogous results for the splitting criterion – Gini impurity was chosen once again in preference to the information gain.

binning	baseline	cross-validation	max_depth
none	0.333	0.456 ± 0.032	10
$(-1), (0), (1), (2), (3, 4), (5–10), (>10)$	0.333	0.510 ± 0.036	14
$(-1, 0, 1), (2, 3, 4), (>4)$	0.693	0.776 ± 0.027	14

Table 5.5: Results of random forest classifier on *train-1*.

binning	baseline	cross-validation	max_depth
none	0.315	0.456 ± 0.014	18
$(-1), (0), (1), (2), (3, 4), (5–10), (>10)$	0.315	0.537 ± 0.018	16
$(-1, 0, 1), (2, 3, 4), (>4)$	0.657	0.784 ± 0.018	18

Table 5.6: Results of random forest classifier on *train-all*.

5.4.3 Decision Tree Regression

The results of the regression experiments with the decision trees are captured in Tables 5.7 and 5.8. As for the parameters: since the only splitting criterion available is mean squared error, the only free parameter was the maximum depth – its value corresponding to the best result is listed in the last column again.

transformation	baseline *RMSE*	cross-validation *RMSE*	baseline R^2	cross-validation R^2	max depth
none	8.12	6.323 ± 3.80	0	0.1578 ± 0.264	3
$ln(y + 2)$	0.8338	0.6016 ± 0.080	0	0.4657 ± 0.054	5

Table 5.7: Results of decision tree regressor on *train-1*.

transformation	baseline *RMSE*	cross-validation *RMSE*	baseline R^2	cross-validation R^2	max depth
none	8.85	7.165 ± 0.43	0	0.3382 ± 0.105	5
$ln(y + 2)$	0.87	0.5949 ± 0.008	0	0.5275 ± 0.024	7

Table 5.8: Results of decision tree regressor on *train-all*.

5.4.4 Tree and Forest Analysis

An example of a decision tree model trained for classification can be seen on the Figure 5.1. For the sake of clarity, this tree has only depth of 3, although in our experiments all our classifier models are deeper. However, the top of our classifiers actually could look very similar, due to the nature of the tree construction algorithm. From our picture, we can see for example that if the given instance is not a noun (i.e. value 0 of feature `m_tagPOSThis=NN`) and the nearest preceding noun or pronoun is further than 1 word, then it has quite high probability of being classified to the leap height value of 0. And if the instance is a possessive reflexive pronoun (`m_tagPOSThis=P8`), then this classification will be almost certain (with a very low probability of an error).

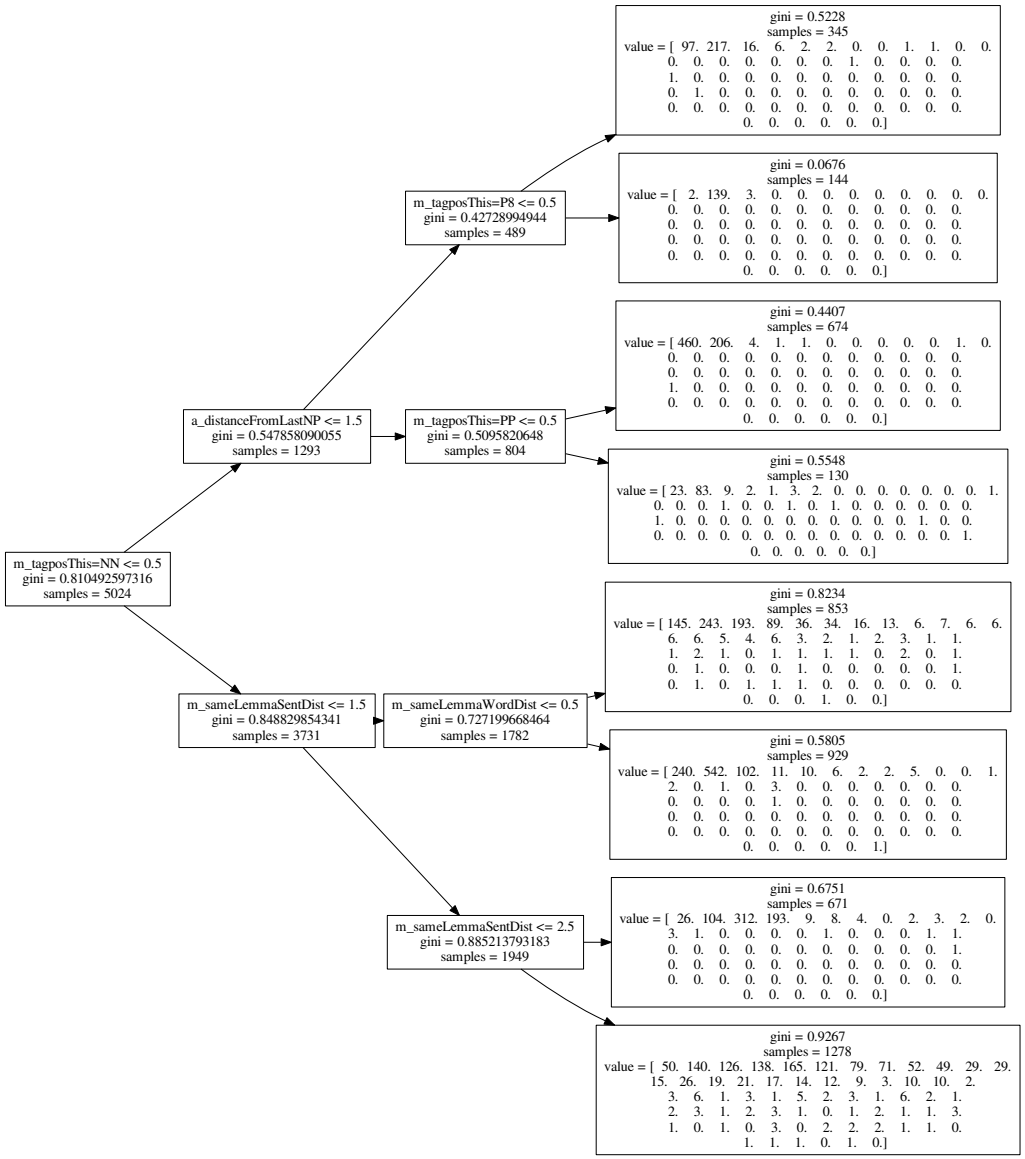


Figure 5.1: Decision tree trained on *train-1* with maximum depth set to 3. Satisfying the condition in a node means going to its upper child. The "value" array in the leaf nodes (on the right) denotes the "buckets" with numbers of training instances with the given target value falling to this leaf. The "buckets" correspond to ordered values of the leap height, i.e. $[-1, 0, 1, 2, 3, ...]$.

5.4.5 Discussion

The results presented above show that to a certain level of success, the leap height variable is learnable from the given features. The results are all well above their respective baselines, and with the exception of the regression task without the logarithmic transformation, this improvement is significant. However, the absolute success rates are not very high, making the predictions not very reliable. The target value transformations introduced here improve the results in each case quite as anticipated, although while losing some amount of expressivity and usability of the predictions.

Variable importance The importance scores of the first 10 features (variables) are listed in Table 5.9, based on one of the trained random forest models. The higher the score, the more important is the variable. The values serve only for relative comparison of the features since their absolute values are dependent on the training set.

rank	variable	importance
1	sentence dist. to last same lemma	0.112
2	word dist. to last same lemma	0.077
3	POS+subPOS tag denotes (any) noun	0.063
4	sentence-position from start of document	0.060
5	word dist. to last agreeing noun	0.051
6	word-position from start of sentence	0.048
7	rel. position to predicate	0.044
8	word-position to end of sentence	0.042
9	sentence-position to end of document	0.042
10	sentence dist. to last agreeing noun	0.033

Table 5.9: Variable importance based on the random forest model trained on *train-all* with the 7-bin binning scheme.

We can see from the comparison that the two dominant decisive features for the leap height classification are the "CR-like" features expressing distance to the last word with the same lemma (complemented by the other two concerning the last agreeing noun, located lower in the table). This seems meaningful as the distance of the coreference antecedent, if recognized properly, is the main source of information for the salience notion, and consequently also for the salience leap height. This direction might be perhaps supported in a future research by adding some more features of this orientation, exploiting more the mechanisms of coreference resolvers (mentioned in Section 5.3.1).

Another predictor which seems relatively strong is the distinction between nouns and pronouns in the surface expression. This agrees with the hypothesis formulated in Section 4.2.3 that the surface expression form is influenced by the previous salience degree. The importance of this feature confirms this claim somehow from the inverse point of view.

Among the features ranked quite high there is also the word position within the sentence (being it the distance from its beginning or to its end, or the position relative to the sentence predicate). This is likely to be attributed to the TFA influence on leap height from its definition, combined with the fact that TFA is tied closely with the word order.

Perhaps an improvement in the results could be achieved by adding more features focusing also on the potential antecedent (hopefully detected by the "CR-like" features), such as the last-same-lemma word. For example trying to use its position in a sentence to approximate its TFA value.

Yet another interesting approach could be to run an automatic coreference resolver on the data and use its results as a foundation for other features (with a certain level of reliability). This would somehow separate the coreference information from the rest of the features, whereas our approach tries to model them altogether.

Regression with ordinal classes Our case of predicting an integer-valued target feature can be seen as a task somewhere between classification and regression – and as such, none of these "pure" methods is actually ideal. This type of problem is sometimes called *ordinal logistic regression* or prediction of *ordinal classes* and is addressed in the context of decision trees for example by Kramer et al. (2001). Their approach suggest slight modifications to the regression tree model to reflect that we are not working with a proper real-valued target, it could be characterized as "clever rounding" at the right time. Unfortunately, the desired modification of the model would be rather complicated in the environment used in our experiments, but any further attempt for improvement is encouraged to examine also this possibility.

Conclusion The main goal of these experiments was to examine the possibility of whether the salience leap height variable can be predicted reliably enough to be used as a feature in further machine learning tasks. The results unfortunately does not confirm this hope directly, the accuracy of the models trained here is not clearly persuasive from this point of view. However, it can be stated that the

results are good enough for the leap height to be used as a side feature or as one of a larger set of features to help improving a further machine learning model.

Furthermore, we believe that these experiments do not exhaust the potential of the machine learnability of the leap height (or salience) value, and perhaps some other methods could bring a significant improvement in such results. While the decision tree(-based) models are invaluable for the analysis, in some tasks they lack the ability to capture dependencies among some groups of features; a problem that another type of model might address more successfully.

6. Document Clustering Experiments

In this chapter, a visual comparison of the document relations was created between the document information contrasting simple word counts against average salience of coreference chains. The aim of this series of experiments is to form an idea of how the salience information could contribute to the information about document relations. We explore the possibility of whether incorporating the salience data alters the relations between documents, possibly revealing some connections or partitioning not visible when working with their vocabulary only.

This visualization is based on representing the document collection as a graph – with documents as the graph nodes and a pairwise document *overlap* defining the graph edges. The contrastive comparison was then made by changing the definition of how the *overlap* is computed. For these experiments, the *train-1* dataset was used, containing a total of 316 documents, 289 of which contains at least one coreference chain. Since the core of the experiments are the pairwise relations between documents, there is no urgent need for a larger dataset, thus this one was chosen to speed up the experiment processing.

6.1 Phase One: List Cutting

The generic idea of this experiment was for each document to list some of its characteristic items in the order of their supposed importance, cut this list off at some point, and then look for matching items in the other documents' lists.

6.1.1 Sorting the Nouns by Counts

One of the most straightforward and yet most frequently used features of extracting key words from a document is the word count. Usually it is complemented by a filter of stop-words, but in our case, when we have information about the word types at our disposal, the simplest way is to work with nouns only.

Document noun-overlap In this sense, the document overlap based on the noun-counts is constructed in the following way:

For each document:

1. extract all the distinct nouns (their lemmata) along with their counts,
2. sort the nouns according to their counts,
3. cut the list at 10% of its length, so that the most frequent nouns remain.

Then, the *noun-overlap* of two *documents* is defined as the number of nouns which are present in both lists, i.e. the size of their intersection.

6.1.2 Sorting the Chains by Average Salience

Average salience, adjusted When looking for an optimal measure for ordering the whole coreference chains in terms of a coarse informative representativeness, the average salience is a natural choice. However, to avoid favoring chains which first occur lately in the document, the simple average must be adjusted to better reflect their inactivity before their first occurence. According to the idea that the items represented by these chains are present in the stock of shared knowledge (but not mentioned yet), their initial course is simulated similarly as if they had been mentioned in the topic of the first sentence. Thus, until their first mention, they undergo a descent by 1 from the value of -1. The general formula for computing the average salience of chain referring to an object x in a given document is then as follows:

$$\begin{aligned} AvgSal(x) :&= \frac{1}{N}\left(\sum_{i=1}^{m-1}(-i) + \sum_{i=m}^{N} dg_x^i\right) \\ &= \frac{1}{N}\left(-\frac{(m-1)m}{2} + \sum_{i=m}^{N} dg_x^i\right) \end{aligned} \tag{6.1}$$

where m is index of the sentence with the first mention of x and N is the total number of sentences in the document.

Document chain-overlap Having defined the average salience, the document overlap based on that measure can be constructed analogously:

For each document:

1. for each coreference chain, extract the list of nouns from the chain along with the average salience of the whole chain,
2. sort the chains according to their average salience,
3. cut the list at 10% of its length, so that the most salient chains remain.

The overlap of two *chains* is defined as the number of members which are present in both two chains, i.e. the size of their intersection. Then, the *chain-overlap* of two *documents* is defined as the sum of overlaps of each chain pair from these documents.

6.1.3 Clustering Visualization

A collection of documents can be viewed as a graph where each node represents one document. Two documents have a common edge iff there is a (non-zero) *overlap* between them, and the weight of this edge equals the size of this overlap. Then, with the help of a commonly used graph visualizing tool (in our case, `pygraphviz`, see Section 3.3.4), the two graphs resulting from the definitions above were drawn for a visual evaluation.[1] The resulting graph visualization of the noun-based overlaps and salience-based overlaps can be seen on Figures 6.1 and 6.2, respectively.

Only the nodes with at least one edge are included in the graph, i.e. documents without an overlap with any other are left out. Furthermore, edges with noun-overlap value of only 1 are omitted from the noun-overlap graph for better clarity. These edges do not significantly influence the graph structure and the visualization would be hardly readable with them.

The sizes of the graph nodes represent relative sizes of the corresponding documents. Also the thickness of the edges reflects the relative size of the document overlap, although these differences are not very large; the chain-overlap values varies from 1 (most common) to 4, for the noun-overlap from 1 to 8 (visible eges only 2 to 8).

6.1.4 Discussion on the Clustering

As we can see, the first graph has no clear smaller clusters, although proximity between some of the documents hints about some share of their vocabulary (with higher word-counts). However, the documents do not form any visually distinguishable groups. Whereas in the second picture, signs of several small clusters can be perceived, suggesting possible topic relations within some document subsets.

Another interesting thing to note is the "real" number of nodes and edges. Although the two graphs have a comparable number of nodes (178 and 127,

[1] The graphviz layout program used was `NEATO` with model `subset`, for more details refer to North (2004).

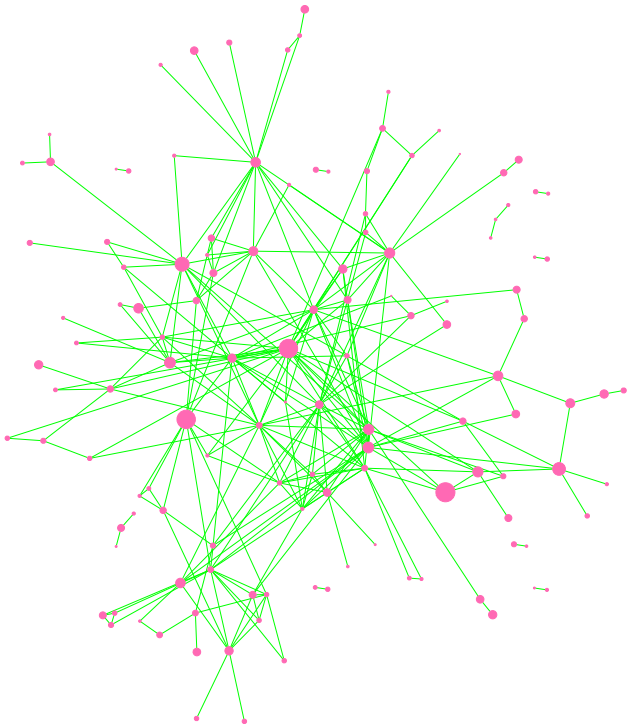


Figure 6.1: Noun-based document overlap with the list-cutting approach.

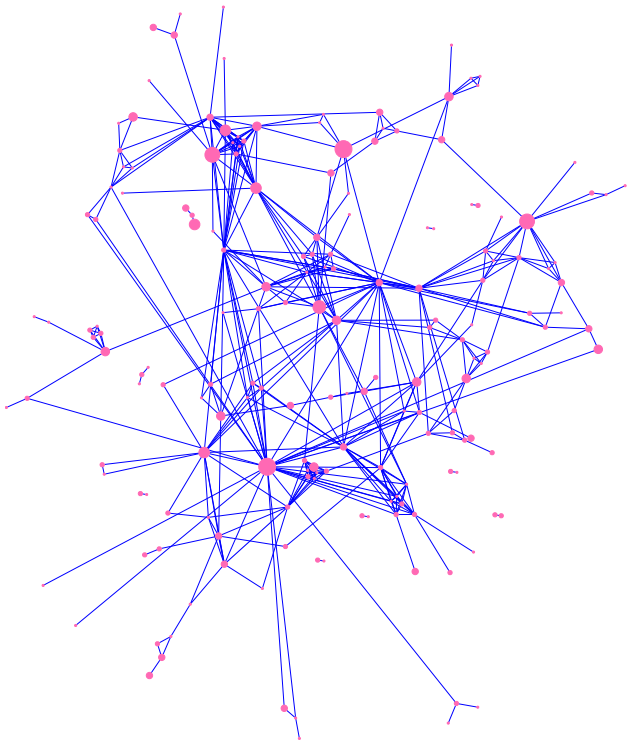


Figure 6.2: Salience-based document overlap with the list-cutting approach.

respectively) and edges (275 and 443, respectively) shown in the pictures, this follows from the omission of the one-noun-overlap edges from the first picture. If these edges were present, the number of the noun-graph edges would rise to 3768 and number of nodes to 275 (because some nodes would have obtained their first overlap, thus entering the picture). The smaller amount of chain-overlaps in the second graph is a direct consequence of the narrower vocabulary when focusing on the coreference members only.

In fact, if the omitted edges in the noun-overlap graph (Figure 6.1) were present, a slight "clustering" would actually emerge. There would be one large group of roughly half of the documents connected together because of sharing a very common noun: "rok" (*"year"*). Also a similar, but smaller group would be formed around the word "Praha" (*"Prague"* – the capital of the Czech Republic). But both these words are really common and can be expected in the documents regardless of their topic, especially when working with journalist texts.

Let us look closer at what chains are shared within the cluster in the upper section of the chain-overlap graph (Figure 6.2), where five or ten documents seem to have a closer connection among them. Their shared word is "strana" (*"(political) party"*), nothing much else. Similar small clusters can be found elsewhere in the picture around words like "svaz" (*"(labor) union"*), "společnost" (*"company"*) or "autor" (*"author"*). These groupings are somehow promising to reflect at least general contents of the documents, but almost all the involved connections are supported by a single-chain overlap only, which does not seem very robust.

The two presented pictures and their analysis revealed that the chain salience can be beneficial to retrieve some more information about document relations (in the sense of their similar topics) than simple word counts. However, limitations of this approach are that it relies on a very narrow data evidence and delivers topics which are too general or vague. The main problem with this list-cutting method is reducing the vocabulary or chain base before computing the pairwise document overlaps, thus resulting in their smaller number. A better approach would be to keep the base for the overlaps as wide as possible, but engage the word counts and salience values in a way that will support the potential "keyness" of the words rather gradually. An attempt for such method is presented in the following section.

6.2 Phase Two: Weighing the Overlaps

We address the deficiencies of the previous approach by retaining the whole list of extracted nouns or chains, but using the word-counts or average salience for *weighing* the overlaps instead. Thus, the definition of the two types of overlaps will be as follows:

Noun-based weighted overlap of two documents D_1, D_2 is a sum of word-counts (c_1 and c_2, respectively to each document) of each noun w present in both documents:

$$NWOverlap(D_1, D_2) := \sum_{w \in D_1 \cap D_2} c_1(w) c_2(w) \tag{6.2}$$

Salience-based weighted overlap of two documents D_1, D_2 as list of coreference chains x is a sum of chain overlaps (according to the definition from the previous approach – see 6.1.2) each multiplied by a weight based on their average salience value:

$$ChWOverlap(D_1, D_2) := \sum_{\substack{x_1 \in D_1, \\ x_2 \in D_2}} |x_1 \cap x_2| \, weight(x_1) weight(x_2) \tag{6.3}$$

where

$$weight(x) := -\frac{1}{AvgSal(x)} \tag{6.4}$$

.

Note that each definition of the overlap results in rather different magnitude of the weights. While the multiplication of the word-counts can yield numbers in orders from units to tens or hundreds, the results from the chain overlap weighing will vary between 0.001 and 0.8. This is not a problem, as only their relative sizes are crucial for the visualization.

6.2.1 Clustering Visualization

The visualization procedure is very similar to the previous phase (see Section 6.1.3), except for the fact that more edges are omitted in both pictures. Since the aim of this experiment was to keep the base for document overlaps as wide as possible, the resulting number of edges grew greatly in both cases and it would render the pictures unreadable to include all of them. Therefore, an arbitrary weight threshold was set in each case in order to keep the graph readable, resulting in

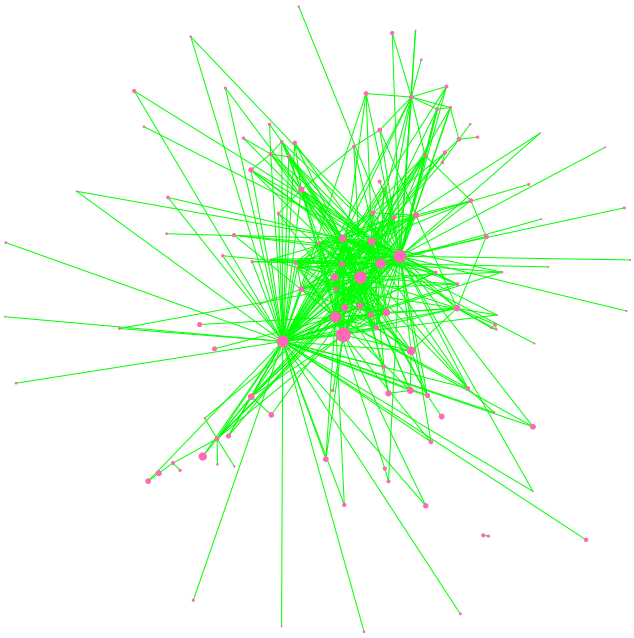


Figure 6.3: Noun-based document overlap with the overlap-weighing approach.

the number of displayed nodes being between 100 and 200. Other rules for the visualization are kept the same way as in the previous experiment.

The resulting pictures are shown in Figures 6.3 and 6.4.

6.2.2 Discussion on the Clustering

The first thing to notice when looking at the pictures is again a quite different density of the two graphs. Indeed, with a comparable number of nodes (124 and 163, respectively), the number of edges differs more significantly; 453 v. 271 (in terms of average degree of the graph nodes, this is ca. 7.3 v. 3.3). The reason is basically the same as in the previous approach, it is the narrower vocabulary when focusing on the coreference members only.

As for cluster groupings in the pictures, the situation has changed a little, but no evident improvement is visible. No clear clusters can be seen in the first picture, although the large group in the middle seems to have something in common. More thorough examination reveals that there is again quite a small set of nouns involved strongly in the overlaps. In this case, it is especially the word "dítě" (*"child"*) or again "rok" (*"year"*). (The latter is also responsible for the long "beams" from one of the nodes slightly on the left. This effect is caused by its multiple, but relatively weak bindings with documents which would have no other sufficient overlap with anything else.) While some of the other words seem

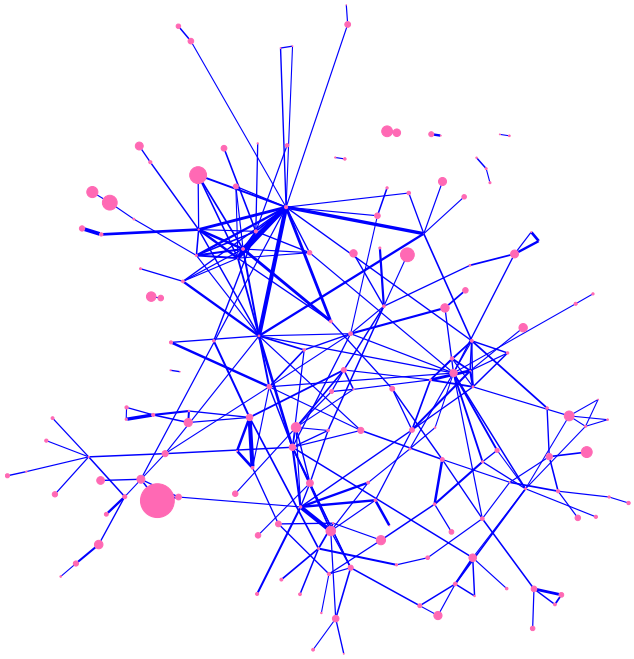


Figure 6.4: Salience-based document overlap with the overlap-weighing approach.

to point better to a possible common topic (such as combination of "dítě"– *"child"* and "výchova"– *"upbringing"*), these bindings are still too confusing to establish a certain topic group within our set of documents.

In the second picture, some smaller clusters are again somehow recognizable, but perhaps not as clearly as in the previous approach. The denser group in the left upper part is formed mainly again around the repeating word "rok" (*"year"*), whereas the small group in the middle right is brought together by the word "strana" (*"(political) party"*). When enquiring deeper, few other nodes are tied by a promising set of words "banka", "smlouva", "úrok" (*"bank"*, *"contract"*, *"interest"*), unfortunately this node group is quite small and hardly recognizable from the visualization (it is located in the lower left part of the picture).

Whereas the weighing approach seemed to bring some little improvement to the word-count overlap sensitivity, the problems of the chain-overlap detected in the previous approach were not quite solved. Some of the observations show somehow promising relations, but the salience information does not seem to be fully exploited by these methods. Furthermore, the visualization of the structures were not really enhanced by this change of approach.

6.3 Conclusion and Outlook

The goal of this experiment was to get a basic idea of how the salience information could alter the results of a computational analysis of a document collection. Having no "correct" results of real proximity or topic relationships between documents, any numeric evaluation would be hard to accomplish. Therefore, a simple graph visualization in an attempt for a visual clustering was performed in the frame of two experiments and evaluated by a manual examination of its outcome. Each of these experiments compared a visualization based on information from word-counts of documents (serving as a rough "baseline") with a salience-based visualization. The two experiments differed in the way of using the information in the data processing before feeding into graph; the first one was based on filtering, the second one on weighing.

The presented results suggest that the salience information might be beneficial in examining the properties of document relations, in comparison to a simple word-count statistic. However, this benefit did not prove in our experiments to be very strong, at least not enough to reveal some conclusive thematic partitioning of the document collection.

For similar experiments in a future research, it will perhaps be helpful to find a better distinction between simply common words and words which are actually useful for identifying the topic of a document. A vast inspiration can be found in the popular field of information retrieval, e.g. engaging concepts like (TF/)IDF (*term frequency/inverse document frequency*) in some way during the weighing phase (or even for the filtering approach).

Our experiment also presented only one way of visualizing the possible relations based on some kind of distance between graph nodes. This way was chosen primarily by its accessibility, relative flexibility and speed. Although perhaps visually appealing, this approach has inherent limitations, especially when trying to display in two-dimensional space something which does not have an explicit location. Other types of visualization techniques might prove better for this task – for example hierarchical clustering (based on mutual distance of nodes) displayed by a dendrogram.[2]

It should also be noted that from a practical point of view, comparing salience with the word-count information as a "baseline" is not quite a fair challenge. Bearing in mind that the salience data inherently contains far more information, a better option for comparison perhaps would be defining the document dis-

[2] A small experiment in this area has already been conducted by the time of finishing this thesis, but was not eventually included due to a need for additional tuning and verifications.

tance exploiting coreference chains (only). This would ensure that the benefits of salience would be proper, above anything we could have in our hands so far.[3]

[3]Some data for this comparison have already been prepared, but in this context of not so persuasive comparison even with the word counts, we eventually decided not to include them in the thesis.

Conclusion

We have presented a reproduction and a data-oriented analysis of the salience algorithm formulated earlier, along with visualizing its results in a larger scale and confirming some of the hypotheses behind the salience notion. This was achieved using the data of the Prague Dependency Treebank 3.0, especially its annotation of the coreference relations and the topic-focus articulation. A brief experiment with the bridging anaphora annotation data was conducted in an attempt for broadening the coverage of the salience models, but deeper investigation in this field remains to further research.

The visualization procedure suggested earlier was made more robust and automatized to allow larger amount of documents to be processed. Also it was extended with procedures which makes the results human-accessible even in this scale.

Another key points of this work were attempts to interpret the output of the salience procedure, the salience graphs. A notion of *salience leaps* and their *height* was introduced and used to confirm the hypothesis about the importance of salience in the decisions about the morpho-syntactic form of the referent.

The salience *leap height* value was then used for exploring the possibility of predicting the salience degrees automatically. Decision trees and random forests models were used in these experiments, using various features extracted from surface, morphological and syntactic layer of annotation and adding some features inspired by the mechanisms of coreference resolution. The direct evaluation results of these experiments were quite positive, but the accuracy of the approach used was assessed to be probably not reliable enough for a further usage of the target value as a main feature in a subsequent model, perhaps feasible as one of complementary ones. Nevertheless, also other valuable information was acquired from the model analysis, such as structure and importance of the variables for the decisions. Further possibilities for an improvement in this task were proposed in the conclusion of this analysis.

Finally, two experiments in the area of document visualization were performed to estimate a possible contribution of the salience information in a field of document processing. By displaying a document collection in a graph representation with a contrasting definition of node distances, these experiments compared the salience-based information with a simple word-count statistics. The evaluation was based solely on the visual comparison with a hope for the salience data to unveil a previously invisible structure among the documents. Although no defini-

tive clusters were perceived, the salience-based visualization seemed to perform a little more promising than its counterpart.

Our hope is that this work may stimulate a future research in this promising area as there are plenty of door opening in this field. The last chapter hopefully might serve as a first small step of the salience data into the machine learning territory, the analysis and discussion revealing some of its further possibilities.

Bibliography

Barzilay, R. and Lapata, M. (2008). Modeling Local Coherence: An Entity-Based Approach. In *Computational Linguistics*, vol. 34(1):pp. 1–34.

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). *Prague Dependency Treebank 3.0*. `http://ufal.mff.cuni.cz/pdt3.0/`.

Bejček, E., Panevová, J., Popelka, J., Smejkalová, L., Straňák, P., Ševčíková, M., Štěpánek, J., Toman, J., Žabokrtský, Z., and Hajič, J. (2011). *Prague Dependency Treebank 2.5*. `http://ufal.mff.cuni.cz/pdt2.5/`.

Breiman, L. (2001). Random forests. In *Machine learning*, vol. 45(1):pp. 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A. ISBN 978-0412048418.

Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In *Subject and Topic* (edited by C. N. Li), pp. 25–55. Academic Press, Cambridge, MA, USA.

Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its applications to software engineering. In *Software: Practice and Experience*, vol. 30(11):pp. 1203–1233. `http://www.graphviz.org`.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. In *Computational Linguistics*, vol. 21(2):pp. 203–225. ISSN 0891-2017.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive Status and the Form of Referring Expressions in Discourse. In *Language*, vol. 69(2):pp. 274–307.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., and Mikulová, M. (2006). *Prague Dependency Treebank 2.0*. `http://ufal.mff.cuni.cz/pdt2.0/`.

Hajičová, E. (1993). *Issues of sentence structure and discourse patterns*. Charles University Press, Prague, Czech Republic.

Hajičová, E. (2003). Contextual boundness and discourse patterns. In *Proceedings of XVII International Congress of Linguists, CD-ROM*, pp. x1–x7. Matfyzpress, MFF UK, Prague, Czech Republic.

Hajičová, E. (2013). Contextual boundness and discourse patterns revisited. In *Discourse Studies*, vol. 15(5):pp. 535–550.

Hajičová, E., Havelka, J., and Veselá, K. (2005). Corpus Evidence of Contextual Boundness and Focus. In *Proceedings of the Corpus Linguistics Conference Series*, pp. 1–9. University of Birmingham, Birmingham, UK.

Hajičová, E., Hladká, B., and Kučová, L. (2006). An Annotated Corpus as a Test Bed for Discourse Structure Analysis. In *Proceedings of the Workshop on Constraints in Discourse Structure Analysis*, pp. 82–89. National University of Ireland.

Hajičová, E. and Vrbová, J. (1982). On the role of the hierarchy of activation in the process of natural language understanding. In *Proceedings of the 9th conference on Computational linguistics*, vol. 1 of *COLING '82*, pp. 107–113. Academia Praha, Prague, Czechoslovakia.

Halliday, M. A. K. (1967). Notes on Transitivity and Theme in English: Part 1. In *Journal of Linguistics*, vol. 3(1):pp. 37–81.

Haviland, S. E. and Clark, H. H. (1974). What's new? Acquiring New information as a process in comprehension. In *Journal of Verbal Learning and Verbal Behavior*, vol. 13:pp. 512–521.

Holan, T. and Žabokrtský, Z. (2006). Combining Czech Dependency Parsers. In *Text, Speech and Dialogue* (edited by P. Sojka, I. Kopeček, and K. Pala), vol. 4188 of *Lecture Notes in Computer Science*, pp. 95–102. Springer Berlin Heidelberg. ISBN 978-3-540-39090-9.

Joshi, A. K. and Weinstein, S. (1981). Control of Inference: Role of Some Aspects of Discourse Structure-Centering. In *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 385–387.

Koo, T., Rush, A. M., Collins, M., Jaakkola, T., and Sontag, D. (2010). Dual Decomposition for Parsing with Non-Projective Head Automata. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1288–1298. Association for Computational Linguistics, Cambridge, MA.

Kramer, S., Widmer, G., Pfahringer, B., and De Groeve, M. (2001). Prediction of Ordinal Classes Using Regression Trees. In *Fundamenta Informaticae*, vol. 47(1-2):pp. 1–13.

Kuno, S. (1972). Functional Sentence Perspective: A Case Study from Japanese and English. In *Linguistic Inquiry*, vol. 3(3):pp. 269–320.

Kučová, L. and Hajičová, E. (2004). Coreferential Relations in the Prague Dependency Treebank. In *Proceedings of the 5th International Conference on Discourse Anaphora and Anaphor Resolution Colloquium*, pp. 97–102.

Nedoluzhko, A. (2011). *Rozšířená textová koreference a asociační anafora. Koncepce anotace českých dat v Pražském závislostním korpusu*. ÚFAL – Institute of Formal and Applied Linguistics, Prague, Czech Republic.

Nedoluzhko, A., Mírovský, J., and Novák, M. (2013). A Coreferentially annotated Corpus and Anaphora Resolution for Czech. In *Computational Linguistics and Intellectual Technologies*, pp. 467–475. ABBYY, Yandex, ABBYY, Moscow, Russia. ISBN 978-1-937284-58-9.

North, S. C. (2004). Drawing graphs with NEATO. In *NEATO User Manual*, vol. 11. `http://www.graphviz.org/pdf/neatoguide.pdf`.

Novák, M. and Žabokrtský, Z. (2011). Resolving noun phrase coreference in czech. In *Anaphora Processing and Applications*, pp. 24–34. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research*, vol. 12:pp. 2825–2830.

Poláková, L., Jínová, P., and Mírovský, J. (2014). Genres in the Prague Discourse Treebank. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 1320–1326. European Language Resources Association, Reykjavík, Iceland. ISBN 978-2-9517408-8-4.

Poláková, L., Jínová, P., Zikánová, Š., Hajičová, E., Mírovský, J., Nedoluzhko, A., Rysová, M., Pavlíková, V., Zdeňková, J., Pergler, J., and Ocelák, R. (2012). *Prague Discourse Treebank 1.0*. `http://ufal.mff.cuni.cz/pdit/`.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In *Syntax and semantics* (edited by P. Cole), vol. 14, pp. 223–255. Academic Press, New York.

Quinlan, J. R. (1986). Induction of decision trees. In *Machine Learning*, vol. 1(1):pp. 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. ISBN 1-55860-238-0.

Sauper, C., Haghighi, A., and Barzilay, R. (2010). Incorporating Content Structure into Text Analysis Applications. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pp. 377–387. Association for Computational Linguistics, Stroudsburg, PA, USA.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. In *Computational linguistics*, vol. 27(4):pp. 521–544.

Spoustová, D. j. (2008). Combining Statistical and Rule-Based Approaches to Morphological Tagging of Czech Texts. In *Prague Bulletin of Mathematical Linguistics*, vol. 89:pp. 23–40.

List of Figures

List of Tables

Appendix – CD-ROM Contents

- `data` folder – Contains sample data from the PDT 3.0 corpus (`pdt_sample`). All the files with the `.t.gz` extension in this folder are in the PML format and ready to be used by the `btred` scripts (however, some of the `btred` scripts will modify them, so make sure the data files have a write permission and you have a backup of them before running those scripts). For viewing the data files, use the TrEd viewer/editor (with the 'bridging' and 'pdt30' extensions installed), available at `http://ufal.mff.cuni.cz/tred/`.
  The subfolder `salience-sample-svg` contains a sample result of the salience procedures, an HTML complex readily browsable from index.html.

- `scripts` folder – All the non-trivial script files used in this work; `btred`, `Perl`, `Python`, `R` files, `bash` scripts. Most of them require to be run on a Linux machine, `btred` scripts require the `btred` application to be installed. The `scripts-readme.txt` file provides the overview of the script files along with a brief information about their functionality and usage.

- `vacl-dipl_thesis.pdf` file – This work in pdf format.

- `readme.txt` file – General information about contents of the CD-ROM.