

List of figures

- Figure 1** An 8-faced dice: Three yes / no – decisions to determine a state of affairs — **6**
- Figure 2** *Noise, equivocation and total amount of information in a message* — **7**
- Figure 3** TCM with general context which includes local contexts — **14**
- Figure 4** Flow of information in a *Garden Path* sentence — **16**
- Figure 5** Uniform information density in Slovenian and Latvian — **17**
- Figure 6** Dependency structure for the sentence *Ich esse einen Keks* ‘I eat a cookie’ — **18**
- Figure 7** Dependency structures of two sentences — **18**
- Figure 8** Distribution of information and UID in eight languages — **19**
- Figure 9** Ascending lexical information content in stacks intensifiers according to the principle of *Uniform Information Density* (left) and violation of this principle (right) — **20**
- Figure 10** Dependency structure of the sentence *Hans eats cookies* with the target word *cookies* and the probabilities *prefprob* and *parse_target* — **21**
- Figure 11** Dependency parse trees of an example sentence and calculation of *prefprob* and *parse_target* — **23**
- Figure 12** Intensional contexts: Information of a signal differs from its meaning — **32**
- Figure 13** Set theoretic representation of digital information and semantic content of propositions — **33**
- Figure 14** Different levels of information — **34**
- Figure 15** Green arrows represent synonym-relations, red arrows hypernym relations, blue arrows location based-relations, violet arrows meronym / holonym relations, brown arrows time based-relations and black arrows co-occurrences relations — **38**
- Figure 16** Fictitious hierarchical structure of a word ontology with verbs — **39**
- Figure 17** Fragment of a graph from *WordNet* — **40**
- Figure 18** Meaning space as a $M(odels) \times P(ropositions)$ matrix — **42**
- Figure 19** Meaning space of the two propositions as truth table — **43**
- Figure 20** Correlations between different types of surprisal of words and semantic entropy reduction — **45**
- Figure 21** The number of extra bits of a text, given a set of background texts, calculated by Pointwise Kullback-Leibler divergence — **46**
- Figure 22** Collocations $v_1 \dots v_j$ in a collocation window, i.e., context window of the target word — **48**
- Figure 23** SVD: Matrix decomposition — **51**
- Figure 24** Maximum informativeness of a semantic unit σ as a definite integral in the interval $[0,1]$ — **54**
- Figure 25** Fuzzy sets of intensifiers — **60**
- Figure 26** Two stacks of intensifiers in expected linear order (left) and in unexpected linear order (right) — **62**
- Figure 27** The correlation between length and surprisal of intensifiers — **64**
- Figure 28** The correlation between amount of surprisal and expressivity of intensifiers — **65**

- Figure 29** The correlation between amount of surprisal and expressivity for some selected intensifiers — **65**
- Figure 30** Transition probabilities of intensifiers as representation of use value — **66**
- Figure 31** Differing numbers of possible transitions with Dutch intensifiers *zo* ‘so’, *echt* ‘really’, *fucking* ‘fucking’ and *lekker* lit. ‘delicious’ — **73**
- Figure 32** IC_{TRANS} increases with the number of transitions — **74**
- Figure 33** Transition probabilities for *echt* ‘really’ — **74**
- Figure 34** A strong negative correlation between IC_{TRANS} and IC_{LOCAL} — **77**
- Figure 35** Negative correlation IC_{TRANS} and IC_{LOCAL} in German Twitter data (blue), and combined data from Twitter and Claudi (2006) (yellow) — **77**
- Figure 36** Negative correlation between IIC_{TRANS} and IC_{LOCAL} in the Dutch data — **79**
- Figure 37** Percentage of each intensifier’s occurrences (pvariants) that are lengthened or capitalised by intensifiers syllable length — **80**
- Figure 38** Density plot of two UID_{LOCAL} distributions: predicted order and violations — **83**
- Figure 39** Difference of flow of information in two verses from the German Luther bible and in a stack of intensifiers — **85**
- Figure 40** IC_{LOCAL} : almost identical increasing flow of information derived from unigrams and bigrams in stacks of intensifiers — **86**
- Figure 41** IC_{TRANS} : decreasing flow of information — **87**
- Figure 42** Path of establishment of intensifiers from high expressivity (bottom right) to high utility (top left) — **88**
- Figure 43** Derivation of the intensifiers-adjective-stack *echt übelst doppelt lecker* ‘really worst double yummy’ — **89**
- Figure 44** Strength of intensifiers as exponent in the membership function — **90**
- Figure 45** Three-dimensional separation of data points by the kernel function and hyperplane — **93**
- Figure 46** Performance of surprisal features in the classification tasks — **94**
- Figure 47** Flow of the analysis in Bizzoni und Lapshinova-Koltunski (2021) — **96**
- Figure 48** Word graph on the above text about President Obama’s visit — **99**
- Figure 49** Workflow of the keyword extraction in the studies of Philipp et al. (2021; 2022) — **102**
- Figure 50** Graphical model of LDA in Hong (2018) based on Blei et al. (2003) — **103**
- Figure 51** Workflow of keyword extraction with a bidirectional GRU in Philipp et al. (2022) — **105**
- Figure 52** Accuracy $a1 - a5$, precision, recall and F1 of NER — **105**
- Figure 53** Results of the TCM — **107**
- Figure 54** Performances of different models in the study of Philipp et al. (2022) — **108**
- Figure 55** Two probabilistic parse (sub) trees with emitted terminal symbols as base of information from the parser model — **109**
- Figure 56** Results of the parser model compared to other models in the study of Kölbl et al. (2020) — **110**
- Figure 57** The term-sentence matrix **U** — **112**
- Figure 58** Results of the study of Philipp et al. (2022): SI from LDA is roughly quite as suitable for keyword extraction as SI from LSA — **113**
- Figure 59** Workflow for determining and iteratively evaluating surprisal and the FoI from complex contexts — **115**

- Figure 60** The size of the text variants with the number of paragraphs, tokens and unique lemmas — **121**
- Figure 61** Mann-Whitney-U-test for mean differences in UIDs and pvalues — **122**
- Figure 62** Comparison of means in information density and extent of scatter of density values in original texts and the texts with fake glosses — **123**
- Figure 63** Information density of meaningfully / fake enriched texts and original texts — **124**
- Figure 64** Cullen and Frey graph for ALL LESK GLOSSES with fake and meaningful enrichment's and original text — **125**
- Figure 65** Coding asymmetries in Mandarin — **128**
- Figure 66** *Coding asymmetry (tendency) for imperfective/perfective, atelic/telic verbs* with default forms shaded in grey — **129**
- Figure 67** Decomposition of the Vendlerian aspect classes — **130**
- Figure 68** Decomposed structure of the extended Vendlerian typology — **131**
- Figure 69** Aspectual composition in an example sentence with coercion — **134**
- Figure 70** Morphologic aspectual coercion — **134**
- Figure 71** Workflow of the prediction of aspectual coding — **137**
- Figure 72** Representation of a sentence in the SynTagRus corpus — **139**
- Figure 73** Data, dependent variable, predictor and covariate in a LMM — **141**
- Figure 74** Dependency structure of the sentence *schrecklichen Wikinger essen den sehr leckeren Gurkensalat* 'the terrible Vikings eat the very tasty cucumber salad' — **143**
- Figure 75** Summary of the best LMM — **144**
- Figure 76** Effect of the three fixed effects *FREQ*, *DEF* and *IC* on verb's lengths — **145**
- Figure 77** Interaction of *IC* and *FREQ* as predictors of verbs' lengths — **146**
- Figure 78** Length (character number of verbs) as a function of *DEFAULT* (left) and interaction of *IC* and *FREQ* for Ancient Greek — **147**
- Figure 79** Length (character number of verbs) as a function of *DEFAULT* (left) and interaction of *IC* and *FREQ* for Russian — **148**
- Figure 80** Length (character number of verbs) as a function of *DEFAULT* (left) and interaction of *IC* and *FREQ* for Basque — **148**
- Figure 81** Length (character number of verbs) as a function of *DEFAULT* (left) and interaction of *IC* and *FREQ* for Hindi — **149**
- Figure 82** Length (character number of verbs) as a function of *DEFAULT* (left) and interaction of *IC* and *FREQ* for Marathi — **149**
- Figure 83** Length (character number of verbs) as a function of *DEFAULT* (left) and interaction of *IC* and *FREQ* for Polish — **150**
- Figure 84** Interaction of *IC* and *FREQ* as predictors of length in six languages — **152**
- Figure 85** Prediction of default and non-default aspectual coding by a n-gram model and three TCM models with different number of topics — **153**

