**Research Traineeships Proposal**

**1. Title of the Project**
What's in a fictitious name? Form-meaning correspondences in character names in children's books

**2. Coordinators**

Dr. Sara Van den Bossche (DCU)
Dr. Giovanni Cassani (DCA)

**3. Project Summary**

**Introduction**

*Nomen est omen* definitely counts for children's book characters, whose fictitious names often attest to their nature. For example, an existing word may be used to name a character [1]: in Swedish, "Pippi" is not at all a common name [2], but rather a rarely-used word, denoting "crazy" or "bird" [3]. The first meaning refers to Pippi Longsotkcing's perception as *fremde Kind* [4,5], the second encapsulates her proximity to animals, adding to her outsiderhood [6]. A second possibility is that character names are made up: for instance, even before we discover it refers to a wizard, "Gandalf" may evoke positivity, following the systematic relation between certain sound combinations and positive semantic domains.

The possibility of guessing word meaning from form alone crucially relies on the hypothesis that form-meaning relations are not entirely arbitrary, contrary to the widely-accepted position that form and meaning are uncorrelated [9]. Supporting our hypothesis, recent studies in cognitive science highlight the occurrence of widespread systematicity in form-meaning relations [10,11], particularly in the vocabulary children acquire early in development [12,13]. We therefore investigate the occurrence and role of systematic form-meaning mappings by considering character names in children's books.

Names are particularly interesting as they refer to individual people, making form-meaning systematicity inherently ill-defined: should we expect all namesakes to have similar traits? However, names are also strongly conventionalized, e.g. some are primarily used for men, others for women. Previous work has shown that participants successfully rely on statistical form-meaning regularities to distinguish predominantly female from predominantly male names [14,15,16,17]. We build on this research to hypothesize that readers can form expectations about several characters' features directly from their name.

Our aim is twofold: first, we aim to scrutinize how meaning-bearing and made-up character names in children's stories are used, the meaning their form suggests, and how readers interpret them; second, we aim to investigate whether we can model these intuitions computationally and explain them relying on the language experience of people. Our main impetus for charting form-meaning correspondences is the opportunity to trace (1) normativity and privileging, and/or (2) insidious discrimination and prejudice [18,19]. Moreover, with this work we aim to investigate to what extent language patterns and sound-meaning correspondences are reliable to infer semantic connotations of character names, meaning-bearing or made-up.

Children's books are particularly fruitful objects of study to probe our theories. Firstly, they deserve scrutiny in light of theories positing a central role of form-meaning systematicity in lexical development, to establish form-meaning mappings for scaffolding language development [12,13]. Secondly, children acquire significant knowledge about the world through reading [23]. Names, in particular, are a potential strategy to characterize and connote figures even before they are described.

## Research Questions

In the light of evidence on form-meaning systematicity and patterns of language use and their importance in language learning, we aim to address the following research questions concerning the general hypothesis that character names in children's books convey consistent intuitions about the characters themselves, whether made-up or meaning-bearing:

- *Do made-up and meaning-bearing character names evoke consistent semantic intuitions about aspects such as valence, age, gender, and ethnicity?*
- *Is there a difference in consistency between made-up and meaning-bearing names, e.g. in the sense that made-up names evoke less consistent semantic intuitions than meaning-bearing names or the other way around?*
- *Is it possible to systematically scrutinize the role of language experience and form-meaning systematic mappings in explaining intuitions about characters' features evoked by names?* In other words, can we build a computational model which approximates people's intuitions about meaning-bearing characters' names by relying on everyday language usage to build semantic representations and about made-up characters' names by relying on systematic form-meaning correspondences in the vocabulary? What kind of information extracted from language does this model have to rely upon?

## Methods

To answer these questions, we combine corpus data, computational linguistics, behavioral experiments, and cognitive criticism. We rely on a sample of books where characters' features and narrative arc have been coded and the most interesting dimensions for further analysis have been identified. We then use distributional semantic methods to capture lexical meaning as a function of co-occurring words and represent meaning using numerical vectors [24,25]. Thusly, we build comparable representations for character names, based on their literary representations, and for thousands of existing words, based on existing large-scale corpora of child-directed language. We analyze whether names align with characters by comparing words with similar semantic representations. In order to get a semantic representation for made-up names, for which it is impossible to leverage co-occurrence patterns, we make use of novel techniques that induce a distributional semantic representation from form alone, exploiting statistical regularities in form-meaning mappings [26,27].

We tie in with implicit bias testing [28] and behavioral studies to explore how names influence people's intuitions about characters. We solicit *unconscious* intuitions through speeded, alternative-forced-choice (AFC) tasks where participants pick the property that best suits a name. Moreover, we collect *conscious* judgments using gliding scales, having participants make their thought process explicit. Importantly, in light of our ideology-critical focus, we do not rely on dichotomies only, but allow for continuums as well. Scales are analyzed quantitatively, open answers qualitatively. Finally, we analyze whether the semantic relations captured in distributional semantic spaces correlate with the behavioral responses, to answer our final research question.

## Trainee Contribution

The trainee will contribute to the entire project and acquire several skills in digital humanities. Learning goals and acquired skills include:

- processing corpus data using NLP tools;
- applying distributional semantics methods;
- contributing to the set-up of crowd-sourcing experiments;
- implementing machine learning methods to derive meaning representations from word form
- performing quantitative data analysis.

A trainee would be immensely valuable to perform more computational simulations and more in-depth analyses of the relation between human intuitions and language experience, and to investigate different characters' features and whether they are encoded in names.

To conclude, this project aims to study whether fictitious, made-up names in children's books convey characters' features, and influence readers' expectations about characters. To address these questions, we combine computational linguistics and cognitive literary theory, using corpora and behavioral paradigms to collect people's un/conscious intuitions about the meaning character names evokes.

## References:

[1] https://www.britannica.com/art/aptronym.

[2] https://www.namn.nu/namn/pippi/.

[3] https://svenska.se/tre/?sok=pippi&pz=1.

[4] Surmatz, A. (2005). *Pippi Långstrump als Paradigma: Die deutsche Rezeption Astrid Lindgrens und ihr internationaler Kontext.* Tübingen/Basel: A. Francke.

[5] Lexe, H. (2003). *Pippi, Pan und Potter: Zur Motivkonstellation in den Klassikern der Kinderliteratur.* Vienna: Ed. Praesens.

[6] Neutelaers, A. & Van den Bossche, S. (2020). Den andra i Astrid Lindgrens Pippiböcker. En postkolonial läsning av skiftande identiteter, kroppsliga egenskaper, och annangörandeprocesser. *Barnboken*, *43*(1) (accepted).

[7] Rudd, D. (2004). The Froebellious child in Catherine Sinclair's Holiday House. *The Lion and the Unicorn*, *28*(1), 53-69. http://muse.jhu.edu/journals/lion_and_the_unicorn/v028/28.1rudd.pdf

[8] Van den Bossche, Sara. (2011). 'The queerer, the better': Een carnavaleske interpretatie van Pippi Langkous. *Literatuur zonder leeftijd*, *25*(84), 87-110. https://www.dbnl.org/tekst/_lit004201101_01/_lit004201101_01_0008.php

[9] De Saussure, F. (1916). *Course in general linguistics*. New York, NY. McGraw-Hill.

[10] Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603 - 615. doi: 10.1016/j.tics.2015.07.013

[11] Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia - a window into perception, thought and language. *Journal of Consciousness Studies*, *8*(12), 3-34.

[12] Imai, M., Miyazaki, M., Yeung, H. H., Hidaka, S., Kantartzis, K., Okada, H., & Kita, S. (2015). Sound symbolism facilitates word learning in 14-month-olds. *PLoS ONE*, *10*(2), 1-17. doi: 10.1371/journal.pone.0116494

[13] Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Science*, *9*(3), 316-22. doi: 10.1111/j.1467-7687.2006.00495.x

[14] Cai, Z.G., Zhao, N. The sound of gender: inferring the gender of names in a foreign language. *Journal of Cultural Cognitive Science* 3, 63–73 (2019). https://doi.org/10.1007/s41809-019-00028-2

[15] Cassidy, K. W., Kelly, M. H., & Sharoni, L. A. J. (1999). Inferring gender from name phonology. *Journal of Experimental Psychology: General*, *128*(3), 362.

[16] Cutler, A., McQueen, J., & Robinson, K. (1990). Elizabeth and John: Sound patterns of men's and women's names. *Journal of Linguistics*, *26*(2), 471-482.

[17] Wright, S. K. (2006). Phonological cues influence sex decisions about novel names. *Psychological Reports*, *99*(2), 315-321.

[18] Nel, P. (2017). *Was the Cat in the Hat Black?: The Hidden Racism of Children's Literature, and Why We Need Diverse Books*. New York: Oxford University Press.

[19] Bernstein, R. (2011). *Racial Innocence: Performing American Childhood from Slavery to Civil Rights*. New York and London: New York University Press.

[20] Bottigheimer, R.B. (2008). Murdering mothers in Bible stories and fairy tales in Germany, 1550-1900. *Women and Death: Representations of Female Victims and Perpetrators in German Culture 1500-2000,* edited by H. Fronius & A. Linton: Camden House, 28-42.

[21] Van den Bossche, S. (2017.) Wat je niet kraakt, maakt je sterker: Roma en Sinti in de hedendaagse Nederlandstalige jeugdliteratuur. *Literatuur zonder leeftijd*, *31*(103), 67-94.

[22] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635-E3644.

[23] Nikolajeva, M. (2014). *Reading for Learning: Cognitive Approaches to Children's Literature.* Amsterdam/Philadelphia: John Benjamins.

[24] Firth, J. R. (1957). *Papers in Linguistics*, 1934-1951. Oxford, UK: Oxford University Press.

[25] Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211-240. doi: 10.1037/0033-295X.104.2.211

[26] Baayen, R. H., Chuang, Y.-Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, *13*(2), 230-268. doi: 10.1075/ml.18010.baa

[27] Cassani, G., Chuang, Y. Y., & Baayen, R. H. (2020). On the semantics of nonwords and their lexical category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(4):621-637. doi: 10.1037/xlm0000747. Epub 2019 Jul 18.

[28] https://implicit.harvard.edu/implicit/takeatest.html.

## 4. Project Timeline

| Task | Months | Deliverable | Supervisor |
|---|---|---|---|
| Corpus creation | 1-2 | A machine-readable corpus of children's books | Cassani Van Den Bossche |
| Set-up of crowd-sourcing experiments | 3 | A set of tasks and stimuli to be administered to participants in the crowd-sourcing experiment | Cassani Van den Bossche |
| Distributional semantic models | 4 | A trained distributional semantic model containing distributional semantic representations for character names and everyday words from the target corpus and a general corpus of child-directed language | Cassani |
| Computational simulations | 5-7 | A code base which yields form-based distributional semantic representations for characters and everyday words | Cassani |
| Data analysis | 7-9 | A statistical analysis about the relation between characters' properties that can be extracted from semantic spaces and people's intuitions evoked by names only | Cassani Van Den Bossche |
| Conclusions | 10 | A report synthesizing the conclusions of the project | Van Den Bossche Cassani |

## 5. Research Trainee Profile

The trainee will focus on the following aspects of the project:

- sourcing corpus data;

- analyze the corpus using NLP tools to build lexical semantic representations for character names and everyday words;
- contribute to setting up crowd-sourcing experiments to collect human intuitions about character names and their likeliest connotations;
- implement machine learning methods to map form to meaning and derive form-based lexical semantic representations;
- analyze behavioral and simulation data to verify whether simulation data approximate behavioral data.

Candidates for this role should exhibit:

- knowledge of programming tools;
- some degree of familiarity with computational linguistics
- familiarity with simple machine learning and statistical modeling techniques
- an observable interest in crowd-sourcing experiments.
- curiosity for literary analyses and corpus research
- preliminary knowledge about psycholinguistic research

Ideally, candidates should be enrolled in a Master or Research Master program, although students in the third year of the bachelor may also apply.

In order to apply, please send your CV and a short motivation statement (max 300 words) to g.cassani@uvt.nl and s.vandenbossche@uvt.nl, having care to indicate why this project interests you and how you think you could best contribute to its success.