Characters, names and reference

Johnsen, Lars
yoonsen@gmail.com
National Library of Norway

Kåsen, Andre
Andre.kasen@nb.no
National Library of Norway

In this presentation we go through some preliminary steps towards the goal of analyzing Norwegian literary texts and map them to a discourse representation. Such a representation makes texts available for a narrative analysis, like extracting properties and relationships between characters, for example along the lines of Piper (2017, 2016).

A character is modeled as its textual occurrences: names, pronominal references and appellations. This will provide the basis for an approximation to discourse models (e.g. Kamp and Reyle (1993)) and Piper (2016). For example, a character “John” in the text “John drinks water, the guy was thirsty, and he is happy”, the words “John”, "guy" and “he” form a chain (“John”,"guy", “he”), which we take to represent the character named John. However, in this presentation we will limit ourselves to proper names and pronouns. The character will in turn be connected to the predicate, drinks, thirsty and happy.

The features we report on here is the inventory of pronouns in Norwegian literature. How long are sequences of pronouns, and how many are there in Norwegian novels? We split the pronouns up into several categories.

Among the referential pronouns are the first, second and third persons, which in Norwegian are: first jeg, second du, third han (masc), hun (fem), det (neut), den (genderless) as well as place indicators: dit and der (there). Additionally there are case variants like possessive forms and accusative as well as unexpressed pronouns like subjects of infinitives. We keep the relative pronouns out from the counts, but will try to accommodate infinitives. We do not take into consideration diegetic level, like indirect speech.

The actual counts for Norwegian are done using the digital resources from the Norwegian National Library as well as the tools from its Digital humanities lab (see digital resources). There are around 600 000 book texts digitally available. For counting word occurences we select a subset equipped with Dewey decimal number series 800 (general fiction), written between 1950 and 2022 with language code ‘nob (Norwegian bokmål)’, which contains approximately 100 000 individual books . In the table below there are two columns, one for all pronouns, and one for the third person gendered pronouns (han, hun), where the numbers indicate percentages of the total:

Table 1:

all pronouns (3rd, 1st, 2nd, singual plural)	The two gendered han, hun + accusative and genitive
mean 5.03 std 1.78 min 0.000578 max 20.36	mean 2.52 std 1.30 min 0.000459 max 10.43

The maximum values clearly represents outliers, given the standard deviation og about half the mean.

We illustrate how to compute the ratio of names to pronouns by using the name recognition tool (see resources) which lets us analyze the content of an arbitrary book. The table below has the numbers for two works: Knut Hamsun’s “Markens grøde” and Agnar Mykle’s “Sangen om den røde rubin”.

Table 2:

Hamsun “Markens grøde”	Mykle “Sangen om den røde rubin”
PRON 15915 (pronouns) PROPN 6110 (names)	PRON 20348 (pronouns PROPN 3731 (names)

These two novels show the suspected span between names and pronouns, here ranging from about one third to one sixth.

The construction of reference chains goes through a series of steps.Step 1 provides a parse of each sentence resulting in parts of speech, syntactic relations and names of people. In step 2 we extract all relations of named objects (e.g. protagonist) and also the list of pronouns. A shallow discourse is constructed as a step 2, by linking the pronouns to their potential referents, here taken to be the named objects. Step 3, final step, collects all names with their properties, subj-verb, verb-obj, subj-pred, so that we get a bag of words connected to each character. Then characters are compared using these cooccurrences

Bibliography

DH-lab at Norwegian National Library https://github.com/NationalLibraryOfNorway/DHLAB

NER / **POS-tool** https://beta.nb.no/dhlab/navn-og-steder/

Piper, Andrew 2017 Studying Literary Characters and Character Networks, DH2017.

Kamp, Hans. and Uwe Reyle, 1993, From Discourse to Logic, Dordrecht: Kluwer