SPISS
Naturfaglige artikler av elever i videregående opplæring

Skille mellom teksttyper ved hjelp av ordlengder

Forfatter: Johanne Bratland Tjernshaugen, Vestby videregående skole

Sammendrag

Det ble undersøkt om det er mulig å skille teksttyper fra hverandre, ved å bruke en modell basert på prosentvis fordeling av ord med ulikt antall bokstaver. Ordlengdefordelingen i romanutdrag og sammendrag fra vitenskapelige avhandlinger, ble analysert og sammenlignet ved hjelp av et program skrevet i Python. Programmet ble brukt til å finne gjennomsnittsandelen ord med n bokstaver for hver tekst. Disse dataene ble brukt til å finne standardavviket for hver ordlengde. En modell ble konstruert basert på de fire ordlengdene der differansen mellom gjennomsnittene til teksttypene var størst, og standardavvikene for ordlengdene ikke overlappet hverandre. Dette gjaldt fordelingen av ord på tre, fire, åtte og ni bokstaver. Modellen ble testet på seks tilfeldig valgte tekster, og den klarte å korrekt avgjøre hvilken teksttype alle de seks tekstene tilhørte. Det styrker hypotesen om at det er mulig å bruke ordlengdefordeling til å skille mellom romanutdrag og sammendrag fra vitenskapelige avhandlinger.

Innledning

Studier har vist at sammenhengen mellom sjanger og andelen ord med ulikt antall stavelser, er tydeligere enn sammenhengen mellom forfatter og ordlengder. Kelih, Antić, Grzybek og Stadlober (2005) har undersøkt ordlengder i 190 russiske brev og dikt av tre forskjellige forfattere. Ved hjelp av to variabler, kunne forskerne skille de to sjangrene fra hverandre i 90 % av tilfellene. Ordlengden i tekstene spilte ingen viktig rolle når det gjaldt å fastslå hvilken forfatter som hadde skrevet teksten (Kelih, Antić, Grzybek, & Stadlober, 2005). I en annen studie sammenliknes 80 slovenske tekster, likt fordelt på sjangrene private brev, nyhetsartikler, dikt og matoppskrifter. Hensikten var blant annet å undersøke fire sannsynlighetsmodeller, og finne ut om noen av dem kunne brukes til å skille mellom teksttypene på bakgrunn av ordlengdefordeling. Den ene av disse sannsynlighetsmodellene, kunne brukes for å skille de tre førstnevnte sjangrene. (Antić, Stadlober, Grzybek, & Kelih, 2006) Det er dermed dokumentert at det er mulig å skille mellom ulike sjangre, ved hjelp av fordelingen av andel ord med ulike lengder.

I svært mange tilfeller er korte ord (ord på inntil seks bokstaver) oftere i bruk, og har et mer generelt innhold, enn lange ord (ord på over seks bokstaver). Lengre ord har vanligvis et mer spesialisert innhold, og brukes sjeldnere. (Holthe, 2007)

Av dette følger hypotesen for prosjektet:

> *Prosentvis fordeling av ord med ulikt antall bokstaver, ordlengdefordeling, kan brukes til å skille mellom romanutdrag og sammendrag fra vitenskapelige avhandlinger.*

Metode

For å teste hypotesen, ble ordlengdefordelingen i romanutdrag og sammendrag fra vitenskapelige arbeider sammenliknet.

Tekstene ble kjørt gjennom et program skrevet i Python, som fant antall ord med *n* bokstaver, og videre hvor mange prosent av ordene som var *n* bokstaver lange. Verdiene ble overført til Excel og analysert der. For å analysere tekstene i Python, ble følgende tegn og deler av tekstene fjernet:

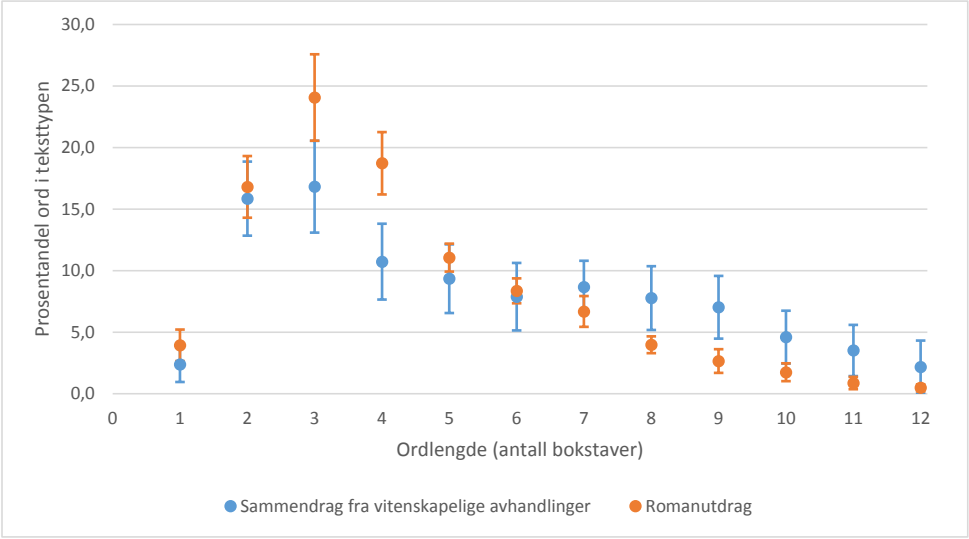
Deler av teksten	Tegn
Overskrifter	. " $
Oppgavemarkeringer: "Task 1", a) o.l.	- ; !
Sidetall, topptekst og bunntekst	, : &
Bilder og figurer	() [] ?
Kildeliste	% ‘
Kildehenvisning i teksten	1 2 3 4 5 6 7 8 9 0

Figur 1: Fjernede deler og tegn

Teksttypene romanutdrag og sammendrag fra vitenskapelige avhandlinger ble valgt, fordi det faglige innholdet i teksttypene er forskjellig. Derfor ville det antakeligvis være forskjell i ordlengdefordelingen. Dessuten er disse to teksttypene lett tilgjengelige. Alle tekstene som ble valgt ut, var engelske. For å lage en modell for romanutdrag, ble de første 2000 ordene i ti romaner fra forskjellige sjangre analysert. 28 sammendrag fra vitenskapelige arbeider ble valgt ut fra tre forskjellige fagfelt: naturvitenskap, humaniora og samfunnsvitenskap, slik at det ble omtrent 2000 ord fra hvert av fagfeltene.

Romanutdragene og sammendragene fra vitenskapelige avhandlinger, ble sammenlignet med hensyn til gjennomsnittlig ordlengdefordeling og standardavvik. En modell for å skille mellom de to teksttypene, ble utviklet basert på de fire ordlengdene der differansen mellom gjennomsnittene var størst, og standardavvikene ikke overlappet hverandre. Modellen ble testet på seks tilfeldig valgte tekster, tre av hver teksttype. Disse var ikke med i tekstgrunnlaget for modellen. Fra romanene ble i overkant av de første 1000 ordene brukt. De ukjente tekstene ble analysert, ved å sjekke om prosentverdiene til de respektive ordlengdene lå innenfor standardavviket til romanene eller sammendragene. Ifølge modellen var den ukjente teksten av den teksttypen som den traff innenfor standardavviket flest ganger. Dersom teksten ikke traff innenfor standardavviket til noen av teksttypene på enkelte ordlengder, ble disse ordlengdene regnet som nøytrale. Modellen avgjorde da teksttypen basert på de resterende ordlengdene.

Resultater



Figur 2: Gjennomsnittlig prosentandel ord på én til tolv bokstaver med standardavvik, i sammendrag fra vitenskapelige avhandlinger og romanutdrag

Punktene representerer gjennomsnittet for teksttypen, mens de vertikale linjene representerer standardavviket. I tre- og fire-, åtte-, ni- og elleve-bokstavers-ord, er det ikke overlapping mellom standardavvikene til de to teksttypene. Differansen mellom gjennomsnittene er størst i fordelingen av ord på henholdsvis tre, fire, ni og åtte bokstaver. Ord på over tolv bokstaver er ikke med, fordi både gjennomsnittsandelen ord på over tolv bokstaver og differansen mellom gjennomsnittene, var svært lave.

Tabell 1: Modell for å skille mellom sammendrag fra vitenskapelige avhandlinger og romanutdrag

	Sammendrag fra vitenskapelige avhandlinger	Romanutdrag
Ordlengde	*Intervall*	*Intervall*
3	13,1 % - 20,6 %	20,6 % - 27,6 %
4	7,7 % - 13,8 %	16,2 % - 21,3 %
8	5,2 % - 10,4 %	3,3 % - 4,7 %
9	4,5 % - 9,6 %	1,7 % - 3,6 %

Modellen er basert på fordelingen av ord på tre, fire, åtte og ni bokstaver, fordi disse ordlengdene har størst differanse i gjennomsnitt, og standardavvikene overlapper ikke hverandre (se *Figur 2*).

Tabell 2: En test av modellen i Tabell 1

Ordlengde	Sammendrag 1	Sammendrag 2	Sammendrag 3	Roman 1	Roman 2	Roman 3
3	15,8 %	14,7 %	20,0 %	25,4 %	34,0 %	26,3 %
	S	S	S	R	N	R
4	10,0 %	10,0 %	16,1 %	19,9 %	22,1 %	22,4 %
	S	S	N	R	N	N
8	5,8 %	16,2 %	6,7 %	2,2 %	3,6 %	2,7 %
	S	N	S	N	R	N
9	6,7 %	6,8 %	4,4 %	2,3 %	1,2 %	2,1 %
	S	S	N	R	N	R
Konklusjon:	Sammendrag	Sammendrag	Sammendrag	Roman	Roman	Roman

Et tilfeldig utvalg av tre sammendrag fra vitenskapelige avhandlinger og tre romanutdrag, som ikke var med i grunnlaget for modellen, ble analysert. Kodene under prosentverdiene for ordlengdene, forteller om teksten var innenfor intervallet til romanutdragene (R), sammendragene fra de vitenskapelige avhandlingene (S), eller ingen av dem (N). Konklusjonen er basert på hvilken av kategoriene romanutdrag og sammendrag som den enkelte teksten traff flest ganger. Ingen av tekstene traff innenfor standardavviket til den motsatte teksttypen i noen av de valgte ordlengdene, og alle tekstene traff innenfor sin egen teksttype minst én gang.

Diskusjon

Analysene av romanutdrag og sammendrag fra vitenskapelige avhandlinger, tyder på at ordlengdefordelingen i teksttypene skiller seg fra hverandre. *Figur 2* viser at gjennomsnittet for sammendragene i ord på inntil seks bokstaver, ligger lavere enn gjennomsnittet for romanene. I ord på sju til tolv bokstaver, ligger gjennomsnittsandelen ord for sammendrag over gjennomsnittet for romaner. Det tyder på at romaner bruker en større andel korte ord, og en lavere andel lange ord, enn sammendrag fra vitenskapelige avhandlinger. Differansen mellom gjennomsnittene er størst i henholdsvis ord på tre, fire, ni, åtte, ti og elleve bokstaver, og ligger mellom 2,8 % og 8,6 %. Standardavvikene for ordene med disse lengdene, bortsett fra ti-bokstavers-ordene, overlapper ikke hverandre. Dette tyder på at det er betydelige forskjeller i ordlengdefordelingen i de to teksttypene, noe som potensielt kan brukes til å skille dem fra hverandre. Forskjellen mellom teksttypene henger kanskje sammen med at sammendragene er svært presise oppsummeringer av vitenskapelige arbeider. I slike tekster er det nødvendig med spesialiserte ord, som i mange tilfeller er lange ord (Holthe, 2007).

Standardavvikene for romanutdrag er lavere enn standardavvikene for sammendrag. Dette kan komme av at det var færre romanutdrag enn sammendrag fra vitenskapelige avhandlinger som ble analysert. Dessuten var romanutdragene lenger enn sammendragene. Ett ord har derfor større betydning i sammendragene enn i romanutdragene, noe som kan ha bidratt til større variasjon blant sammendragene. En annen mulighet er at det i utgangspunktet kan være større variasjon i ordlengden i sammendrag fra vitenskapelige avhandlinger enn i romaner. I vitenskapelige tekster er det ofte lange fagbegreper, som i enkelte tilfeller forkortes. Det er ikke nødvendigvis noen standarder for hvordan og hvilke lange ord som skal forkortes, som kan gjøre at variasjonen i ordlengdefordelingen øker.

Det kan være mulig å bruke modellen i *Tabell 1* til å skille mellom teksttypene romanutdrag og sammendrag fra vitenskapelige avhandlinger, siden forskjellene mellom gjennomsnittene i ord på tre, fire, åtte og

ni bokstaver er relativt store, og standardavvikene ikke overlapper hverandre. Modellen klarte å korrekt avgjøre teksttypen til alle de seks tekstene som modellen ble testet på. Sjansen for å gjette riktig teksttype på seks av seks tekster, dersom sannsynligheten for å gjette riktig per test-tekst er 50 %, er gitt ved . Sjansen for dette er altså svært lav, noe som tyder på at modellen faktisk kan skille mellom romanutdrag og sammendrag fra vitenskapelige avhandlinger. Det peker i samme retning som studiene av de slovenske og russiske tekstene, som også konkluderte med at det kan være mulig å skille ulike sjangre fra hverandre på bakgrunn av ordlengdefordelingen (Kelih, Antić, Grzybek, & Stadlober, 2005) (Antić, Stadlober, Grzybek, & Kelih, 2006). En svakhet ved modellen, er at den ble konstruert på grunnlag av et begrenset tekstutvalg, og at den ble testet på relativt få tekster. Det har ikke blitt testet om resultatet ville blitt det samme med andre tekster i grunnlaget, eller med andre test-tekster.

I tekstene som ble testet, traff særlig romanutdragene utenfor standardavvikene til begge teksttypene i modellen. Roman 2 lå for eksempel over romanutdragenes standardavvik for tre og fire bokstaver, og under for ni bokstaver. Det er likevel verdt å merke seg at verdiene i alle tre tilfellene, lå nærmere intervallene for romanutdrag, enn intervallene for sammendragene. Dette gjaldt generelt for de ordlengdene der test-tekstene ikke traff noen av intervallene i modellen. I av de nøytrale ordlengdene, lå verdien nærmest intervallet til den teksttypen som test-teksten tilhørte. Det at mange av test-tekstene hadde ordlengder som var nøytrale, kan tyde på at standardavvikene, og dermed intervallene i modellen, burde vært større. Både for romanutdrag og sammendrag fra vitenskapelige arbeider ville antakeligvis modellen vært mer nøyaktig dersom det hadde blitt brukt flere tekster for å lage grunnlaget. Flere test-tekster ville også vært gunstig for å teste treffsikkerheten til modellen.

I analysen av tekstene kan det ha oppstått ord som ikke eksisterer, fordi analyseprogrammet fjernet tall og en del tegn. Dette gjelder særlig fagbegreper, der to ord adskilt av tegn eller tall kan ha blitt slått sammen til ett langt ord. Tekstene kan også ha brukt tegn som ikke ble fjernet av analyseprogrammet. I slike tilfeller vil de ukjente tegnene regnes som bokstaver, så det kan virke som det er flere eller lenger ord i teksten enn det virkelig er. Det vil kunne påvirke ordlengdefordelingen i tekstene.

Analysen tyder på at romanene brukte en høyere andel korte ord, og en lavere andel lange ord, enn sammendragene fra vitenskapelige avhandlinger. Testingen av modellen for romanutdrag og sammendrag fra vitenskapelige avhandlinger, styrker hypotesen om at det er mulig å skille teksttyper fra hverandre basert på ordlengdefordeling.

Kilder

Antić, G., Stadlober, E., Grzybek, P., & Kelih, E. (2006). *Word Length and Frequency Distributions in Different Text Genres*. Hentet 16. januar 2018 fra http://www.stat.tugraz.at/stadl/papers/anstgrke06.pdf

Holthe, M. A. (2007). *Ordforråd og syntaktisk kompleksitet i argumenterende elevtekster*. Hentet 15. desember 2017 fra https://www.duo.uio.no/bitstream/handle/10852/26792/HoltheHeleoppgaven.pdf?sequence=3&isAllowed=y

Kelih, E., Antić, G., Grzybek, P., & Stadlober, E. (2005). *Classification of Author and/or Genre? The Impact of Word Length*. Hentet 16. januar 2018 fra http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.518.9763&rep=rep1&type=pdf