



The use of vector models in aggregate-level studies of semasiological variation

Dirk Speelman, Kris Heylen, Dirk Geeraerts

QLVL, KU Leuven

Overview

Introduction

The tool: an application of vector models

Case study

Results

Conclusions





Overview

Introduction

The tool: an application of vector models

Case study

Results

Conclusions





Introduction

- This talk is on a new tool for the aggregate level analysis of (dia)lectal variation



Introduction

- This talk is on a new tool for the aggregate level analysis of (dia)lectal variation
- More specifically it is about a descriptive tool for the detection of semasiological variation in the lexicon



Introduction

- This talk is on a new tool for the aggregate level analysis of (dia)lectal variation
- More specifically it is about a descriptive tool for the detection of **semasiological variation in the lexicon**
- However, at this point we **have not yet reached the stage of using the tool** in actual aggregate-level (dia)lectometric research

Introduction

- This talk is on a new tool for the aggregate level analysis of (dia)lectal variation
- More specifically it is about a descriptive tool for the detection of **semasiological variation in the lexicon**
- However, at this point we **have not yet reached the stage of using the tool** in actual aggregate-level (dia)lectometric research
- The case study presented today is part of ongoing efforts to **refine and calibrate the measures used in the tool** and to assess their reliability and validity

Introduction

Dimensions of lexical variation (Geeraerts *et al.*, 1994):

- **onomasiological variation:** given the thing or concept we want to refer to, what are the different words or expressions that can be used? (e.g. given my computer)
 - **conceptual onomasiological variation:** given that we want to refer to something, how do we construe/conceptualize it? (e.g. THING versus DEVICE versus COMPUTER versus LAPTOP versus MACBOOK PRO)
 - **formal onomasiological variation:** given that we (roughly) have a concept in mind (e.g. LAPTOP), which of several (near) synonyms do we use? (e.g. 'notebook' versus 'laptop' versus 'portable')
- **semasiological variation:** given a word or expression, what are the different concepts or referents that it can refer to? (e.g. given 'notebook')



Introduction

Today's perspective:

- In most (dia)lectometric studies on lexical variation, the variant that is investigated is formal onomasiological variation.
- Today, however we'll look at semasiological variation.



Introduction

The phenomena we're interested in:

- We want to use corpus data to look into regional variation
- At the moment, our main interest is in the difference between Belgium and The Netherlands (but hopefully in the future, data will be available that also allow is to look at more fine-grained regional variation; cf. talk Liberman and talk by Grieve)
- We (eventually) want to investigate, aggregating over many words, which overall regional patterns of semasiological variation there are (= preferences for different senses and different idioms)



Introduction

Examples of (future) research questions:

- How much semasiological variation is there?
- In which places (genres, semantic domains, ...) do we find it?
- Is semasiological variation strongest in the same places (genres, semantic domains, ...) as formal onomasiological variation?

Caveat:

- caveat: because obviously semasiological variation is very sensitive to topic bias, our corpora must be as comparable as possible (e.g. senses of 'goal' in political corpus versus corpus on football).

Introduction

Today's questions:

- Do the measures introduced in the new tool yield sensible results in our case study?
- How valid are these measures?
- And how reliable are they?



Overview

Introduction

The tool: an application of vector models

Case study

Results

Conclusions



Tool

The tool reuses and combines existing techniques:

- **Step 1:** first we use **vector space models** (henceforth **VSMs**) to obtain a representation of how a particular word is used in the corpora that represent our lects. Today we will compare data from two corpora. The result is a matrix (our **vector space**) in which for each of the corpora a number of instance (=tokens) of the word are being stored, and in which distances between these tokens represent how different the usage of the words is in these tokens.

Tool

The tool reuses and combines existing techniques:

- **Step 1:** first we use **vector space models** (henceforth **VSMs**) to obtain a representation of how a particular word is used in the corpora that represent our lects. Today we will compare data from two corpora. The result is a matrix (our **vector space**) in which for each of the corpora a number of instance (=tokens) of the word are being stored, and in which distances between these tokens represent how different the usage of the words is in these tokens.
- **Step 2:** then we apply **cluster quality measures** to that vector space in order to assess to which extent the tokens from the different corpora form different clusters (which would indicate a regional semasiological difference).



Tool

The tool reuses and combines existing techniques:

- **Step 3:** next, we use MDS to obtain a 2D simplification of the structure from step 1, that can be visualized. This we call the [reduced vector space](#).

Tool

The tool reuses and combines existing techniques:

- **Step 3:** next, we use MDS to obtain a 2D simplification of the structure from step 1, that can be visualized. This we call the [reduced vector space](#).
- **Step 4:** finally, we also apply the [cluster quality measures](#) from step 2 to the reduced vector space. (So the tool introduces two sets of measures, which we will compare today.)

Type-based vector space models

The most commonly used kind of VSMs are so-called **type-based VSMs**. In these VSMs

- the usage of a particular word in a corpus is summarized in a single **row** in the VSM (which as a whole is a matrix)
- the **columns** represent so-called features of the word (typically several thousands of them). In the VSMs we discuss today these features simply are words that occur in the vicinity of the target word (i.e. of the word represented in that row of the VSM)
- the **cells** express the frequency with which target words co-occurs with certain features, or rather, they contain so-called PMI values that are derived from the raw frequencies. PMI values express the 'attraction' between a target word and a feature.



Type-based vector space models

The rationale then is that, given a large enough corpus, words with similar meanings tend to have similar rows. Therefore **distances between rows** (typically calculated as one minus the positive cosine of the angle between the two vectors) can be used as a **proxy for differences in 'meaning' (or at least 'usage')**.

	<i>home</i>	<i>drink</i>	<i>traffic</i>	<i>wheel</i>	...
car
vehile
coffee
...

Type-based vector space models

This technique is well-established in NLP and Information retrieval, where it is commonly, and successfully, used in applications such as synonymy detection, thesaurus extraction, etc. (even though it must be added that the technique is by no means flawless, and that it does produce a fair amount of noise together with the sensible patterns it detects).

Variation detection with VSMs

Type-based VSMs from different corpora (but with the same features) could be used to detect regional variation. For an approach along these lines, see Peirsman & Speelman, 2009.

	<i>home</i>	<i>drink</i>	<i>traffic</i>	<i>wheel</i>	...
car in US
car in UK
vehile in US
vehile in UK
coffee in US
coffee in UK
...

However, today we'll look at a more fine-grained approach.

Token-based vector space models

In token-based VSMs each instance (=token) of a word in corpus, or at least a representative number of such instances, has its own row.

	?	?	?	?	...
car 1
car 2
car 3
car 4
car 5
car 6
...

But what then are the features?



Token-based vector space models

What are the features?

- If we took the collocates of the token (henceforth Cs), we would have a very sparse matrix.
- Therefore we look for a richer representation of these Cs (richer than just one cell per C)
- What we do, for each of these Cs (typically just a handful per token), is we take their type vector in a type-based VSM; henceforth we call the features of these type vectors the CCs, i.e. the collocates of the collocate of the target word.
- The token vector of the token then becomes the weighted sum of the type vectors of all its Cs, with the weight being the 'attraction' (PMI) between the target word and the C (this PMI is also found in the type VSM)



Token-based vector space models


What we get:

	C_1	C_2	C_3	C_4	...
car 1
car 2
car 3	<i>weighted sum of C_s of car 3</i>				
car 4
car 5
car 6
...

Token-based vector space models

And now we can merge token-based vector spaces from different corpora (provided the same CCs are used):

	CC_1	CC_2	CC_3	CC_4	...
car 1 from US
car 2 from US
car 3 from US	<i>weighted sum of Cs of car 3 from BE</i>				
...
car 1 from UK
car 2 from UK
...

And we can calculate (cosine-based) distances between all token. 

Cluster quality measures

Next we want to assess to which extent the tokens from the same corpus cluster together

We'll use four measures, ...:

- ... that all **operate on the distance matrix** that resulted from step 1
- ... two of which are **global cluster quality measures**: they assess whether globally clusters can be detected that coincide with the different corpora
- ... two of which are **local cluster quality measures**: they assess whether locally in vector space tokens tend to belong to the same corpus



Cluster quality measures

First global measure: **DR**, which stands for mean **distance ratio** (slightly modified version of McClain & Rao, 1975):

- for each item (=token) we calculate A the mean distance from the item to other items from the same corpus
- for each item (=token) we calculate B the mean distance from the item to items from any of the other corpora
- the cluster quality for the item is B/A
- the cluster quality for the complete 'token cloud' (i.e. all items) is the mean cluster quality of the items

Cluster quality measures

Second global measure: **SIL**, which stands for **silhouette width** (Rousseeuw, 1987):

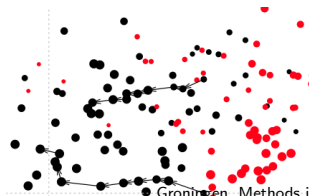
- for each item we calculate to which other cluster (=corpus), apart from its own cluster (=corpus), its mean distance is smallest; we call this the 'neighbouring cluster' of the item
- for each item we calculate A the mean distance from the item to items from its cluster
- for each item we calculate B the mean distance from the item to items from its neighbouring cluster
- the cluster quality for the item is $(B - A) / \max(A, B)$
- the cluster quality for the complete 'token cloud' (i.e. all tokens) is the mean cluster quality of the items



Cluster quality measures

First local measure: SCP, which stands for smallest 'same class path':

- for each item (=token) we calculate the quality of the shortest path of length k that connects it to other items from the same cluster; this quality is the mean 'step quality'; the quality of a step in the path is one divided by the rank of the distance of the next item in the path.
- the cluster quality for the complete 'token cloud' (i.e. all items) is the mean path quality of all items





Cluster quality measures

Second local measure: **KNN**, which stands for *k* nearest neighbours:

- for each item (=token) we calculate the proportion of 'same class items' among its *k* nearest neighbours; that proportion is the cluster quality of that item
- the cluster quality for the complete 'token cloud' (i.e. all items) is the mean cluster quality of the items



Dimension reduction

In the third step we derive a reduced vector space from the original space, using non-metric MDS.

- we want neither the VSM itself nor the cluster quality assessment to be a black box
- therefore we want to be able to visualize the 'token cloud'

The result of this step is

- that we obtain for all items coordinates in 2D space (so that we can visualize the reduced token cloud; see visualization tool)
- from which we can once again derive a distance matrix (by applying Euclidean distances to the coordinates)



Dimension reduction

In the fourth and final step we again calculate cluster quality scores, using the same measures as in the second step, but now starting from the new distance matrix calculated in step 3.

Overview
○

Introduction
○○○○○

Tool
○○○○○○○○○
○○○○○○○

Case study
○○○○○

Results
○○○○○○○○○
○○○○○○○○○
○○○○○○○○○○○
○○○○○○○○○
○○○○○○○○○
○○○○○

Conclusions
○○○

Overview

Introduction

The tool: an application of vector models

Case study

Results

Conclusions



Case study

Case study:

- we built token clouds for 42 words, 21 of which are claimed to display regional semasiological variation (www.wikipedia.org; taaltelefoon.vlaanderen.be), and for 21 of which we found no such claims
- of the former 21 words, 14 are claimed to differ in the possible/popular senses, and 7 are claimed to differ in the expressions/idioms that are often used
- for each of these words we randomly collected 300 tokens from a large Belgian newspaper corpus (LeNC; 1.2 billion words) and 300 tokens from a large Netherlandic newspaper corpus (TwNC; 500 million words), and we merged both sets in one token cloud



Case study

The words:

- word category 'no' (no claims about differences found):
"appel", "auto", "ballon", "bos", "broek", "bureau",
"centrum", "deur", "dier", "fruit", "gebruiker", "heling",
"kamer", "kop", "land", "nacht", "neus", "school", "steun",
"stoel", "verlof"
- word category 'sense' (senses are claimed to differ):
"academicus", "bank", "bolletje", "kleedje", "kous",
"middag", "monitor", "pan", "patat", "poep", "puntje",
"tas", "vlieger", "wagen"
- word category 'expr' (expressions/idioms are claimed to differ): "biecht", "boontje", "geschenk", "mosterd",
"mouw", "straatje", "vijg"



Case study

Case study:

- For each of these words we randomly collected 300 tokens from a large Belgian newspaper corpus (LeNC; 1.2 billion words) and 300 tokens from a large Netherlandic newspaper corpus (TwNC; 500 million words), and we merged both sets in one token cloud.
- We used about 5000 CCs (the intersection of the top 7000 high frequency words in LeNC and TwNC, minus the top 100 high frequency words); the context window used was 4:4.
- We only kept Cs with $LLR > 1$ and $PMI > 1$ in the corpus of the token, and with frequency higher than one in the other corpus; the context window used was 10:10.
- We dropped tokens without suitable Cs (typically keeping about 500 tokens out of the original 600).



Case study

Case study:

- We then calculated the measures DR, SIL, SCP and KNN, both on the original vector space and on the reduced space [in the SCP and KNN measures k was set to 10 and an additional weighting procedure was used that was not explained in this paper]
- Stress in the MDS solutions varied from .15 to .28.
- We repeated the whole procedure five times. So we had five sample sets of each time about 500 tokens for each of the 42 words, and we essentially conducted the case study five times.

Case study

Case study:

Research questions were:

- Do the measures DR, SIL, SCP and KNN on average yield different scores for the categories 'no', 'expr' and 'sense' ? If so, which measures in particular?
- Are the measures valid measures for semasiological variation?
- And are they reliable?



Overview

Introduction

The tool: an application of vector models

Case study

Results

Conclusions



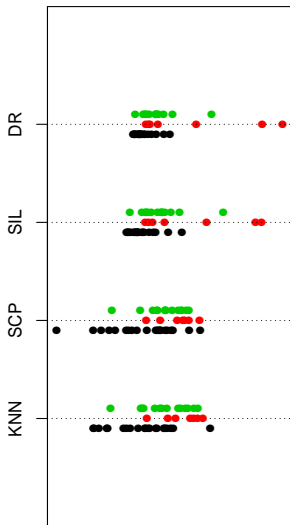
Results

Strip charts

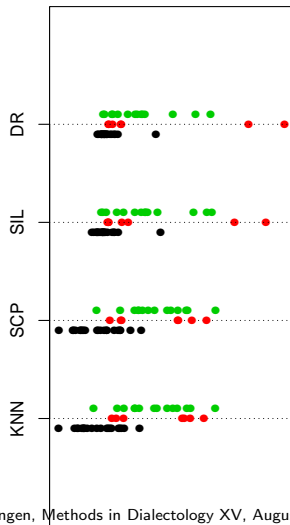
Which patterns do the global measures DM & SIL and the local measures SCP & KNN yield ...

- ... (a) when applied to the original vector space?
- ... (b) when applied to the reduced vector space?

cluster quality
in high-dimensional space



cluster quality
in 2D space

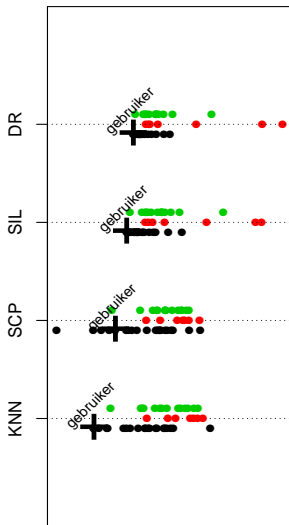


Results

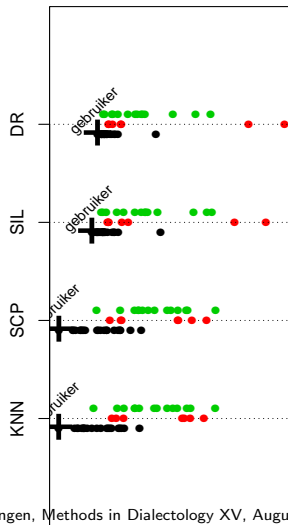
Strip charts compared to eyeballing the token clouds

How does the position on the strip charts compare to the shape of the token clouds in reduced vector space?

cluster quality in high-dimensional space



cluster quality in 2D space

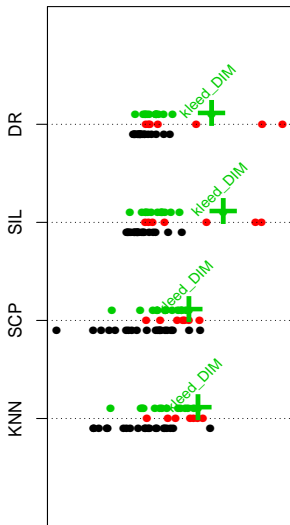


_gebruiker/noun

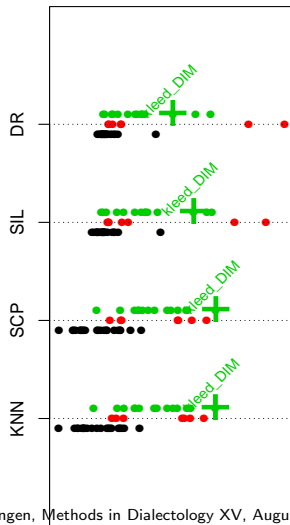
B
N



cluster quality in high-dimensional space

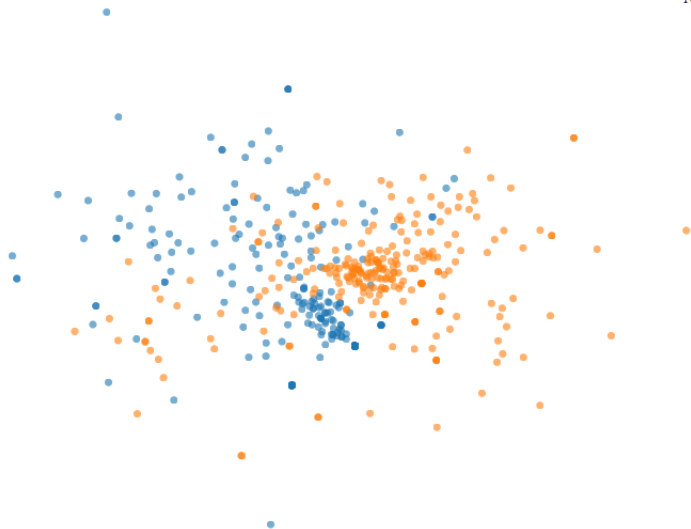


cluster quality in 2D space



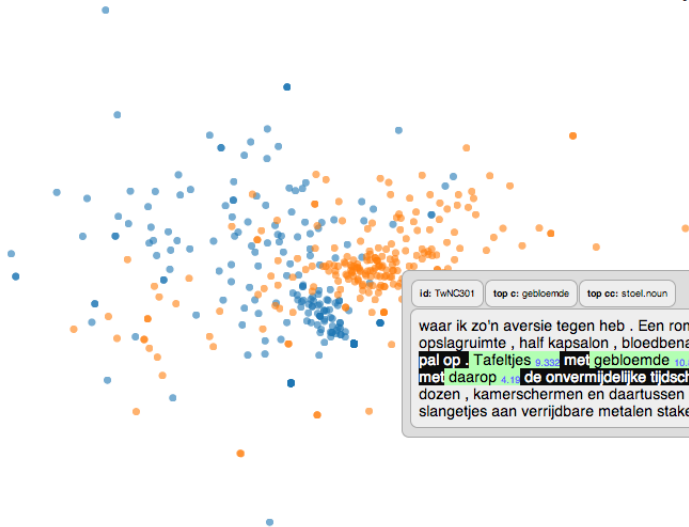
kleed_DIM/noun

B ■
N ■



_kleed_DIM/noun

B 
N 



id: TwNC301

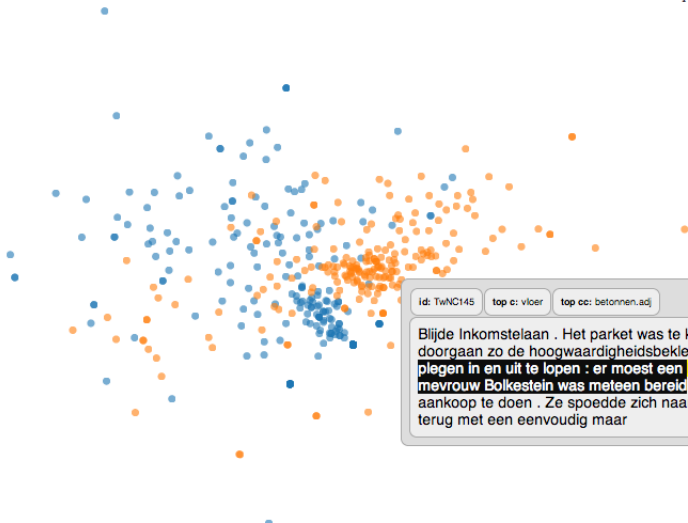
top e: gebloemde

top ee: stoel.noun

waar ik zo'n aversie tegen heb . Een rommelig lokaal , half opslagruimte , half kapsalon , bloedbenauwd ook , **de zon staat er pal op** . Tafeltjes **9.331** met **gebloemde** **10.827** **kleedjes** , **o gezelligheid** , met **daarop** **4.18** **de onvermijdelijke tijdschriften** . verder kartonnen dozen , kamerschermen en daartussen mensen die via een aantal slangetjes aan verrijdbare metalen staketsels zijn verbonden .

_kleed_DIM/noun

B 
N 



Id: TwNC145

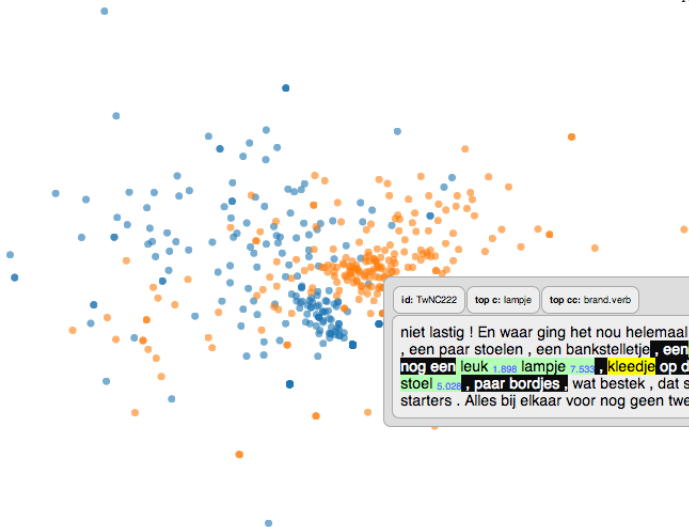
top c: vloer

top cc: betonnen.adj

Blijde Inkomstelaan . Het parket was te kaal ; Frits kon niet doorgaan zo de hoogwaardigheidsbekleders te ontvangen die hier **plegen in en uit te lopen : er moest een kleedje op de vloer** 6,105 . En mevrouw Bolkestein was meteen bereid persoonlijk de nodige aankoop te doen . Ze spoedde zich naar Ikea en keerde weldra terug met een eenvoudig maar

_kleed_DIM/noun

B 
N 



Id: TwNC222

top c: lampje

top cc: brand.verb

niet lastig ! En waar ging het nou helemaal om ? Een keukentafeltje , een paar stoelen , een bankstelletje , een leuk 1,898 lampje 7,533 , nog een leuk 1,898 lampje 7,533 , kleedje op de vloer 6,102 , makkelijke stoel 5,026 , paar bordjes , wat bestek , dat soort handige dingen voor starters . Alles bij elkaar voor nog geen twee ton ! Kleinzielig

_kleed_DIM/noun

B ■
N ■

id: LeNC7624

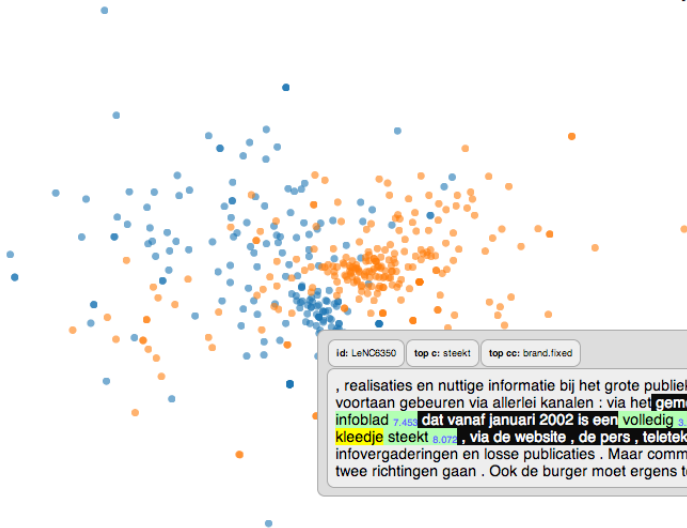
top c: steken

top co: brand.fixed

lang in een slechte staat . Omdat Aquafin langs de Steenweg grote collectorenwerken plant , maakt het gemeentebestuur van de gelegenheid **4.787** gebruik om ook de Steenweg zelf in een nieuw kleedje te steken **8.078** . Tegelijkertijd wordt er met geld van de Vlaamse Landmaatschappij een nieuw fietspad aangelegd . Die werken kaderen in het project Gaverse Scheldemeersen . (PP/DIH)
Nieuwe

_kleed_DIM/noun

B ■
N ■



_kleed_DIM/noun

B
N

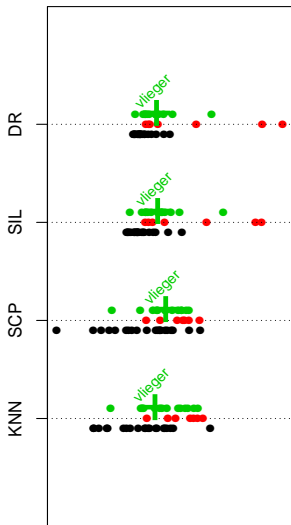
id: LeNC6519

top c: aangemeten

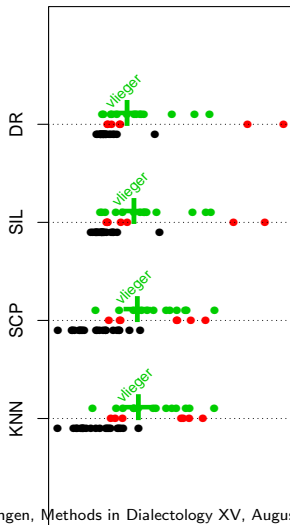
top ce: kostuum.noun

aanduiding van allerlei diensten en instellingen . Ten slotte ontvangen ze ook een speciale editie van de West-Vlaamse Guidogids , een studentenstadsgids . Die kreeg voor de gelegenheid ^{4,701} een Kortrijks kleedje aangemeten ^{9,361} en bevat alle mogelijk informatie die studenten nuttig kan zijn : van het knopen van een das tot en met de strafte cantusstraffen . Achterin steken opnieuw heel wat

cluster quality in high-dimensional space

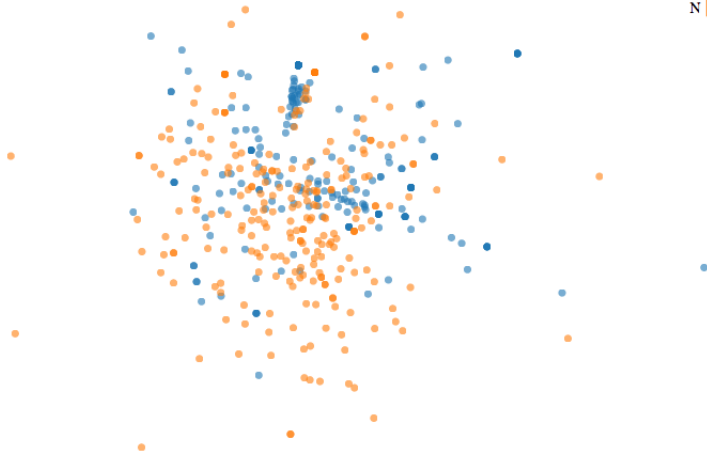


cluster quality in 2D space



vlieger/noun

B 
N 



_vlieger/noun

B ■
N ■

id: TwNC511

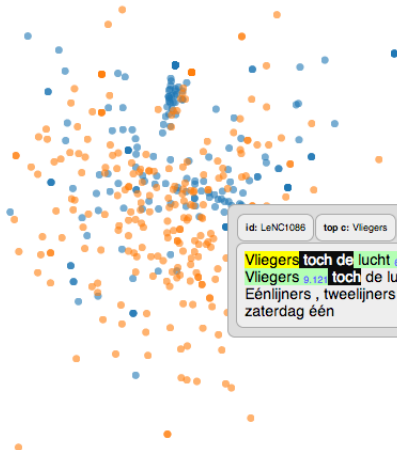
top e: piloot

top cc: piloot.noun

best . Zaken als stijgen en landen spelen nauwelijks een rol . Het toestel reageert nerveus op de besturing en **de automatische piloot** **5.615 stort 5.661 zelfs vaker neer 1.766 dan een onervaren 5.191 vlieger.** **Wat in het oog springt 2.391 zijn de plaatjes.** De beelden van het slagveld zijn van wereldklasse , met bijzonder veel oog voor detail . De aardigste optie is

_vlieger/noun

B 
N 



id: LeNC1086

top o: Vliegers

top co: lucht.noun

Vliegers **toch de lucht** 6,147 **in** ondanks 1,675 gebrek 2,76 **aan** wind 5,085
Vliegers 9,121 **toch** de lucht in ondanks gebrek aan windLOMMEL -
Eénlijners , tweelijners , vierlijners , matrasvliegers ... het was
zaterdag één

_ vlieger/noun

B 
N 

id: LeNC445

top c: gaat

top ee: vlam,noun

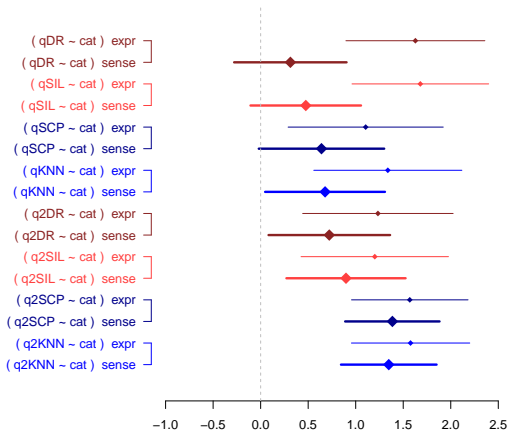
een club geen schadeclaim kan indienen bij een burgerlijke rechtbank als de dader niet bekend is , en dat is **in deze zaak** **vooralnog niet het geval** ^{1,29%} . " Die **vlieger** ^{6,33%} **gaat** ^{1,89%} **dus** voorlopig **niet op voor Westerlo . Coomans** ." Voor een rechtbank van een sportfederatie kan een club wel een klacht indienen tegen de andere club , want

Results

Simple regression analyses

Simple regression analyses as a way of summarizing the information in the strip charts.

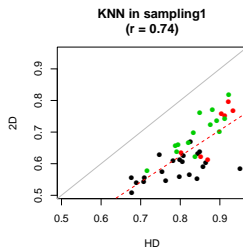
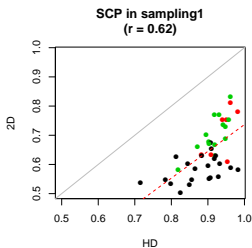
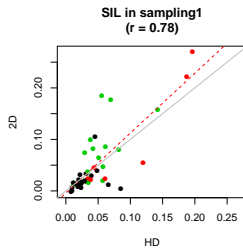
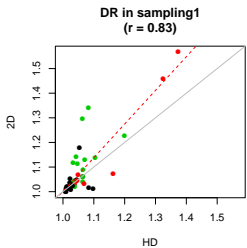
regression analyses q ~ cat sampling1



Results

Correlations between the measures

How do the measures correlate?



Type-based vector space models

Correlations between measures

How do the measures in original space and reduced space compare?

- absolute difference: working with reduced space slightly increases values of global measures DR & SIL
- absolute difference: working with reduced space systematically decreases values of local measures SCP & KNN
- relative difference: for the local measures SCP & KNN working with reduced space tends to relatively decrease the values of `no` and relatively increase the values of `expr` and `sense`.

Type-based vector space models

Validity

Do the measures seem to be a reasonable indication of 'degree of regional semasiological variation' when we manually eyeball the token clouds?

- Thoroughly assessing the validity of the measures will require manually annotating the tokens for several features (manually assigned senses, manual indication of lexical cues, manual or semi-automated identification of collocations, idioms & fixed expressions, ...)

Type-based vector space models

Validity

- Tentative 'impressionistic' first assessment:
 - Words of category *no* that have high scores may indicate a *different treatment/construal of the relevant concept* in newspapers (e.g. 'auto', featuring more prominently in accident contexts in BE, and (to a lesser extent) featuring more prominently in sales and luxury contexts in NL)
 - Words of category *no* that have relatively high scores may indicate the *the presence of several (unanticipated) idioms or fixed expressions that exhibit some regional variation* (e.g. 'neus'; een wassen neus, zijn neus ophalen voor, zijn neus ergens in steken, ...; or 'deur'; met slaande deuren, een voet tussen de deur krijgen, ...; or 'centrum': cultureel centrum; or 'nacht': de nacht van X op Y)

Type-based vector space models

Validity

- Tentative 'impressionistic' first assessment:
 - Words of the categories *expr* and *sense* occasionally have low scores if the relevant expression or sense hardly ever occurs in the genre at hand (e.g. 'bank')

Type-based vector space models

Validity

A general observation is that ...

- There is a lot of variation
- Often this variation is based on differences in the popularity of idioms and fixed expressions
- In future refinements we may try to eliminate these fixed expressions from the data
- However, sometimes there is a sense difference at the basis of the presence of these differences in expressions (e.g. 'kleedje'; in een nieuw kleedje stoppen)



Type-based vector space models

Reliability

How reliable are the measures?

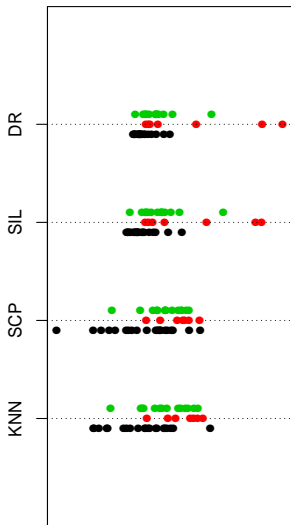
- We'll look at consistency across the five sample sets.

Type-based vector space models

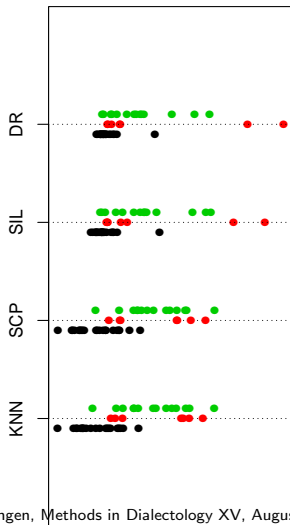
Reliability

- How consistent are the strip charts?

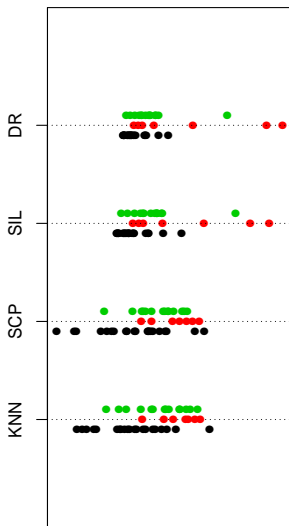
cluster quality in high-dimensional space



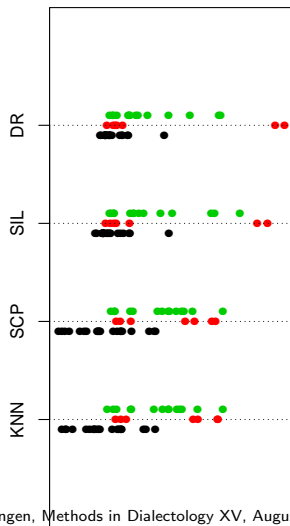
cluster quality in 2D space



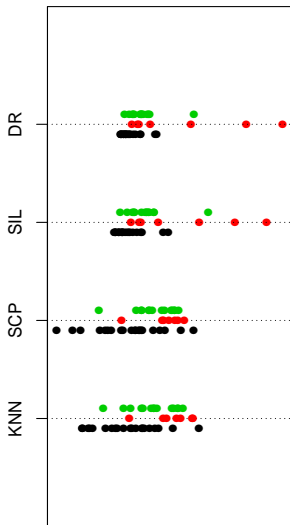
cluster quality
in high-dimensional space



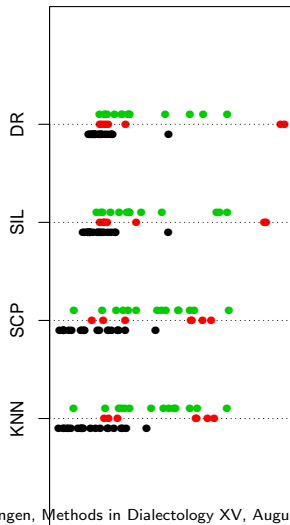
cluster quality
in 2D space



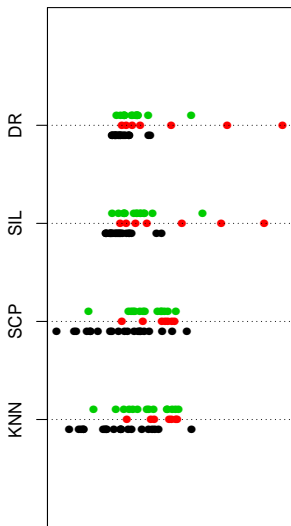
**cluster quality
in high-dimensional space**



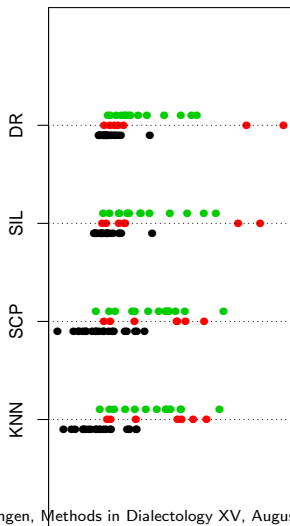
**cluster quality
in 2D space**



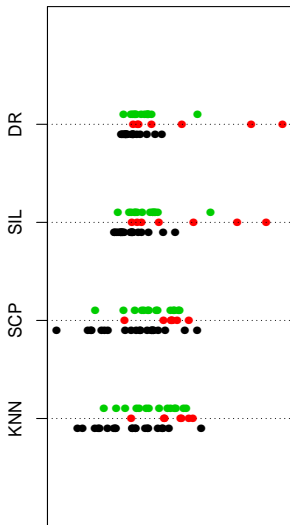
cluster quality
in high-dimensional space



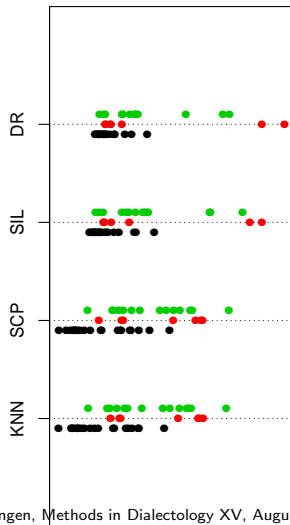
cluster quality
in 2D space



cluster quality
in high-dimensional space



cluster quality
in 2D space

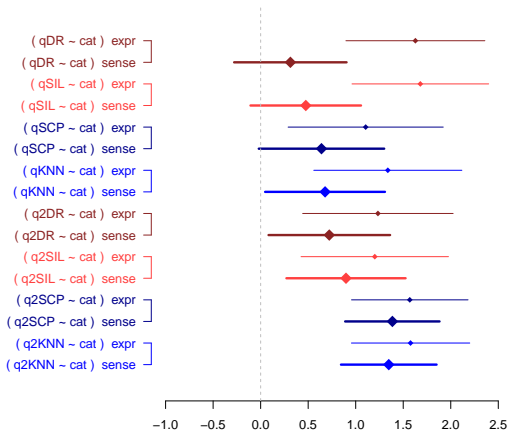


Type-based vector space models

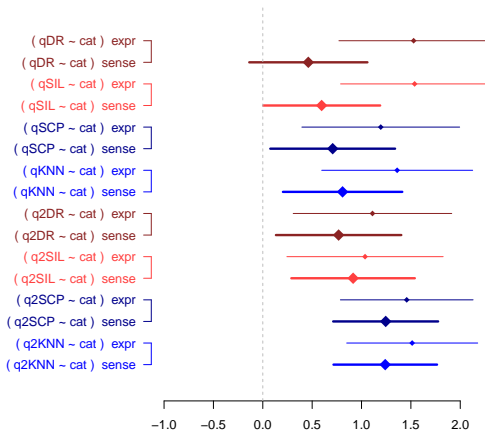
Reliability

- How consistent is the regression output?

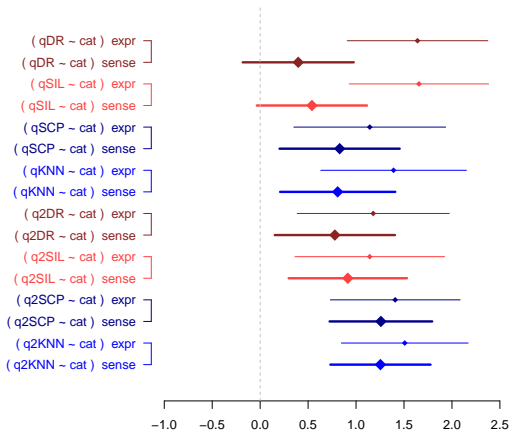
regression analyses q ~ cat sampling1



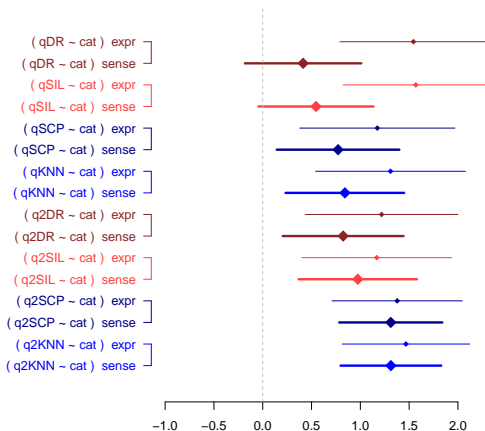
regression analyses q ~ cat sampling2



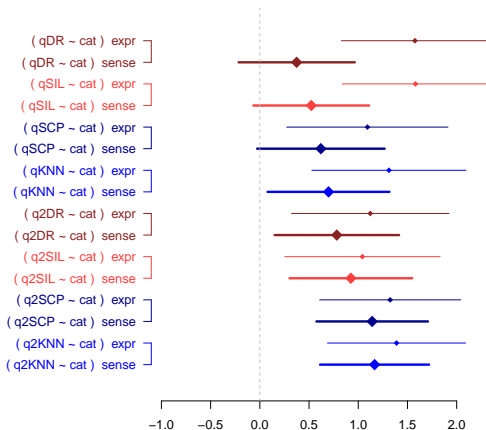
regression analyses q ~ cat sampling3



regression analyses q ~ cat sampling4



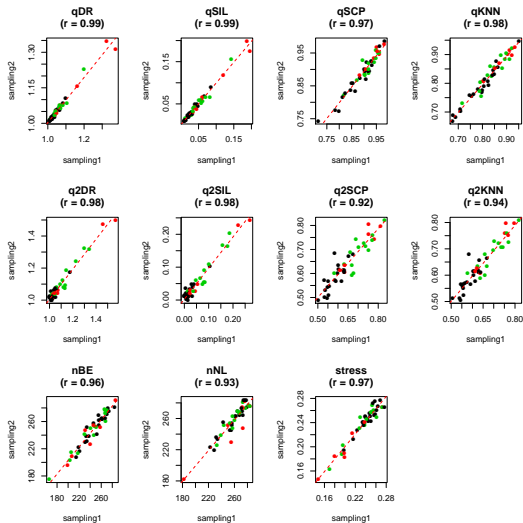
regression analyses q ~ cat sampling5

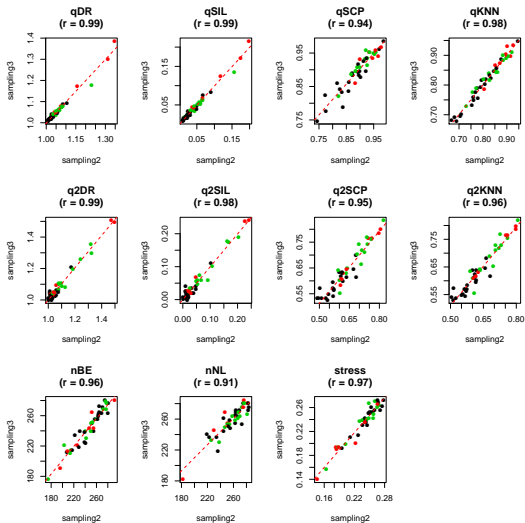


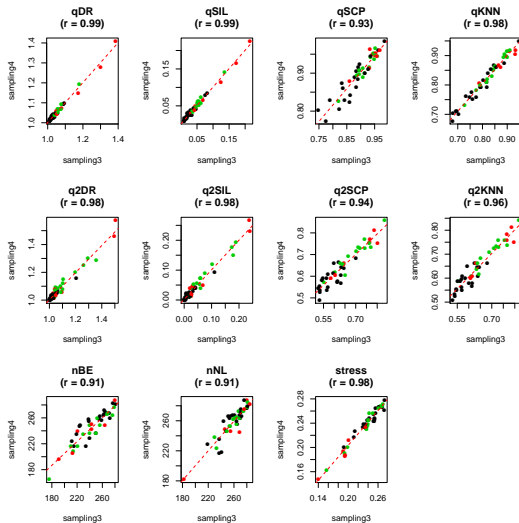
Type-based vector space models

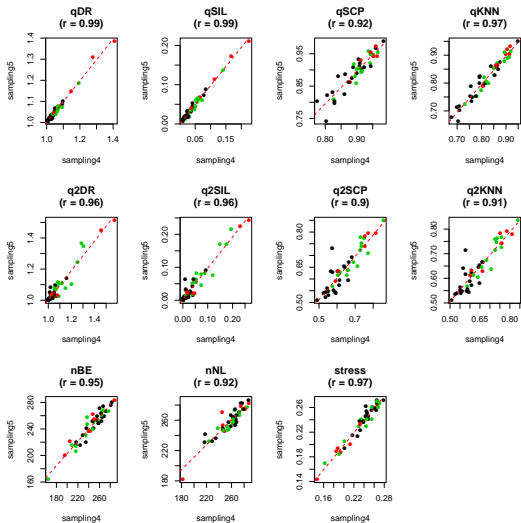
Reliability

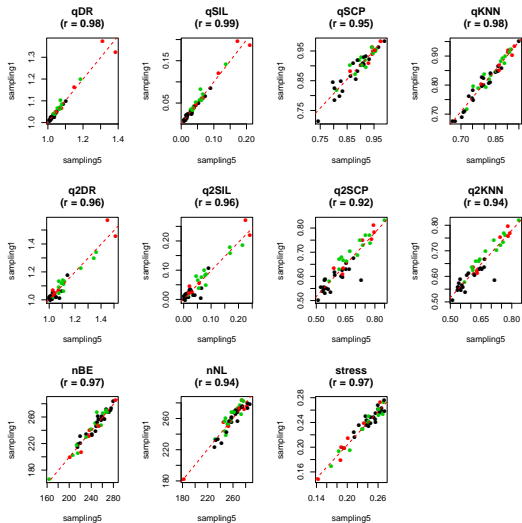
- Another way of looking at the same thing: correlations.







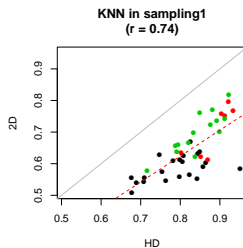
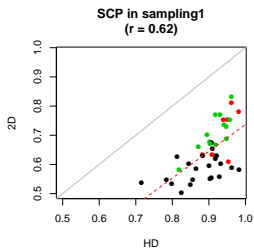
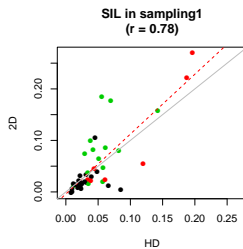
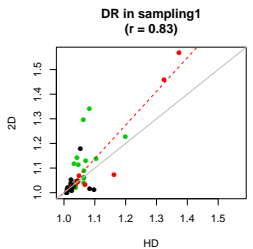


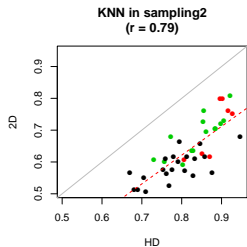
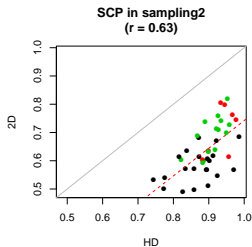
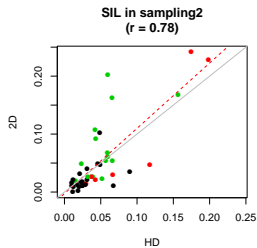
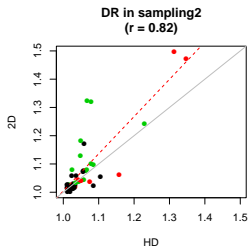


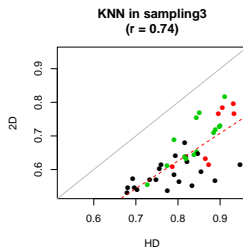
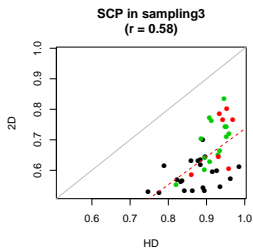
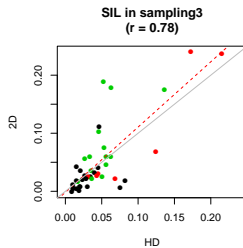
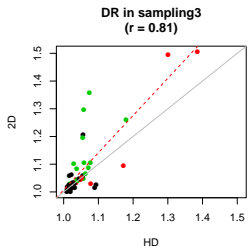
Type-based vector space models

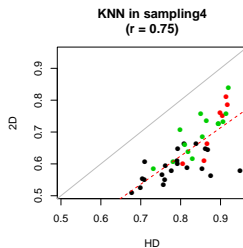
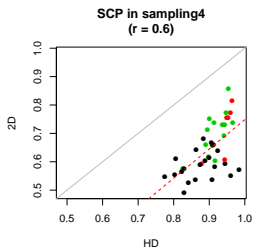
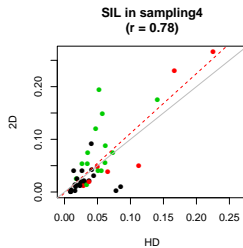
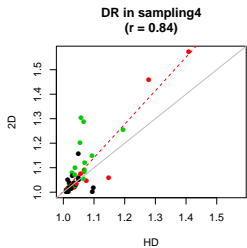
Reliability

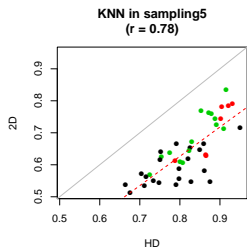
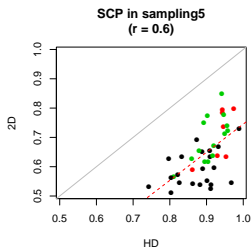
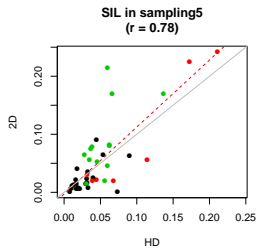
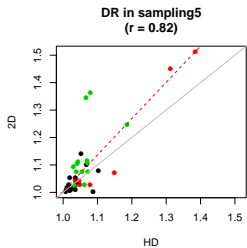
- How consistent is the relation between measures?











Overview

Introduction

The tool: an application of vector models

Case study

Results

Conclusions



Conclusions

Conclusions

Do the measures DR, SIL, SCP and KNN on average yield different scores for the categories 'no', 'expr' and 'sense' ? If so, which measures in particular?

- In general (i.e. for most words) all measures fairly easily distinguish between 'no' and 'expr'
- The distinction between 'no' and 'sense' is harder. Here the local measures SCP and KNN, applied to the reduced vector space, seem to do the best job

Conclusions

Conclusions

Can something be said about the validity of the measures?

- Eyeballing of the token clouds in reduced vector space helps us to find plausible linguistic explanations for the scores in most cases
- This also applies to the cases that deviate from the typical values for the category to which they were assigned (often because the category assignment was based on an underestimation of the variability)
- However, thorough manual annotation of the tokens will be needed to further assess the validity



Conclusions

Conclusions

Can something be said about the reliability of the measures?

- We looked at replicability, and the results across the five sample sets turned out to be remarkably similar

Thank you!

For more information:

`dirk.speelman@arts.kuleuven.be`

`kris.heylen@arts.kuleuven.be`

`dirk.geeraerts@arts.kuleuven.be`

