



Beyond Identity Coreference

Contrasting Indicators of Textual Coherence in English and German

Kerstin Kunz (Heidelberg), Ekaterina Lapshinova-Koltunski, José Martínez-Martínez (Saarbrücken)

> CORBON, San Diego 16/06/16

Background



Research Project

GECCo: German-English Contrasts in Cohesion

supported by the DFG

Project Team:

- Kerstin Kunz
- Ekaterina Lapshinova-Koltunski

- Frich Steiner
- Jose Manuel Martinez
- Katrin Menzel

Acknowledgement: Stefania Degaetano-Ortlieb, Marilisa Amoia

www.gecco.uni-saarland.de

Overview



- Defining Concepts
- Goals and Features
- Methodology
- Results
 - Topic Development
 - Semantic Variation
- Summary



Defining Concepts

Several studies have shown that two of the factors affecting regret are how much one feels personal responsibility for the result and how easy it is to imagine a better alternative. The availability of choice obviously exacerbates both these factors. When you have no options, what can you do? You will feel disappointment, maybe; regret, no. With no options, you just do the best you can. But with many options, the chances increase that a really good one is out there, and you may well feel that you ought to have been able to find it.

Mehreren Studien zufolge wird das Gefühl der Reue zum einen stärker, je mehr man sich für das Resultat persönlich verantwortlich fühlt, und zum anderen, je leichter man sich eine bessere Alternative vorstellen kann. Fin Auswahlangebot verschlimmert offensichtlich beide Faktoren. Was kann man schon groß anstellen, wenn man keine Wahl hat? Vielleicht ist man enttäuscht, aber Reue empfindet man nicht. Wenn es hingegen viele Optionen gibt, wächst das Risiko, dass man meint, eine besonders gute übersehen zu haben, und dies nun bereut.

Several studies have shown that two of the factors affecting regret are how much operations of the personal responsibility Core

result and how easy it

Coreference Studien zufolge wird das Gefühl der Reue zum einen hr man sich für das önlich verantwortlich m anderen, je leichter e. The man sich eine bessere Alternative

imagine a better alternative. The availability of choice obviously exacerbates both these factors When you have no options, what can you do? You will feel disappointment, maybe; regret, no. With no options, you just do the best you can. But with many options, the chances increase that a really good one is out there, and you may well feel that you ought to have been able to

vorstellen kann. Fin Auswahlangebot verschlimmert offensichtlich beide Faktoren. Was kann man schon groß anstellen, wenn man keine Wahl hat? Vielleicht ist man enttäuscht, aber Reue empfindet man nicht. Wenn es hingegen viele Optionen gibt, wächst das Risiko, dass man meint, eine besonders gute übersehen zu haben, und dies hun bereut.

Several studies have shown that two of the factors affecting regret are how much or personal responsibility result and how easy it imagine a better alternative. The availability of choice obviously exacerbates both these factors When you have no options what can you do? You will feel disappointment, maybe; regret, no. With no options you just do the best you can. But with many options, the chances increase that a really good one is out there, and you may well feel that you ought to have been able to find it.

Mehreren Studien zufolge wird das Gefühl der Reue zum einen hr man sich für das Lexical Cohesion ön/ich verantwortlich m anderen, je leichter man sich eine bessere Alternative vorstellen kann. Ein Auswahlangebot verschlimmert offensichtlich beide Faktoren. Was kann man schon groß anstellen, wenm man keine (Wahl) hat? Vielleicht ist man enttäuscht, aber Reue empfindet man nicht. Wenn es hingegen viele Optionen sibt, wächst das Risiko, dass man meint, eine besonders gute übersehen zu haben, und dies nun bereut.

Cohesion



Types of Cohesion (cf. Halliday & Hasan 1976)	Meaning relations
Coreference An option <u>it/ this</u> option	identity
Substitution Many options a good <u>one</u>	Type reference/ comparison
Ellipsis You will feel disappointment. [] Maybe. Many options a good [].	
Comparative Reference One option another/ better option	
Cohesive conjunction X. <u>But</u> / <u>And/ However</u> Y	Logico-semantic relations (addition, contrast, cause,)
Lexical cohesion factors factors	Similarity (repetition, general nouns, meronymy, Synonymy)



Goals and Features

Research Goals



- Variation in coreference and lexical chains as an indicator of
 - topic development
 - topic continuity of individual referents
 - topic continuity within one domain
 - topic variation
 - topic interaction
 - semantic variation
- cross-linguistically: contrasts between English and German, see House (1997, 2015), Hansen-Schirra et al. (2012) and Neumann (2013)
- across registers: between spoken and written registers, see Hundt & Mair (1999), Mair (2006) or Leech et al. (2009)

Goals and Features



topic continuity of individual referents coreference chain length and coref. chain distance

topic continuity within one domain lexical chain length and lexical chain distance and

lexical chain number

topic variation chain length and distance and chain number

topic interaction chain length and chain distance

semantic variation repetition and identity

semantic variation synonymy, meronymy, (co-)hyponymy, hyperony-

my, (co-)meronymy, holonymy, (co-instance), type,

antonymy, synonymy

Example



Differences across registers

FO-INTERVIEW-010

I live in a town called (Reigate). London and the countryside which is quite nice. It takes us about 25 minutes to get to it's a town, it's more of a village. It's quite small. It's very nice actually, it's a nice place to live. And I grew up in a place called Banstead which is fairly close to Reigate.

EO-POPSCI-004

Several studies have shown that two of the (factors) affecting regret are how ... and how easy it is to imagine a better alternative. The availability of choice obviously exacerbates both these factors. When you have no options, what can you do? You will feel disappointment, maybe; regret, no. With no options ... But with many options, the chances increase that (a really good one) is out there, and you may well feel that you ought to have been able to find it.



Methodology

Corpus Data



in tokens

register	EO	GO
ESSAY	27.171	31.407
FICTION	36.996	36.778
INTERVIEW	30.057	35.036
POPSCI	27.055	32.639
TOTAL	121.279	135.860

- subset of GECCo (Lapshinova et al., 2012 and Hansen-Schirra et al., 2012)
- ESSAY and POPSCI: written discourse
- INTERVIEW: spoken discourse
- FICTION: spoken elements in form of dialogues

Annotation

MMAX2 (Müller & Strube, 2006)



1.13.003 /home/katia/saar/gecco/docu/annotationsrichtlinien/LexicalCohesion/LexicalCohesion-GO/GO POPSCI 005.mmax File Settings Display Tools Plugins Info F Show ML Panel nionen , es sind die enzyme , Beginnen wir damit , das wir den Bau dieser für ganz bestimmte chemische warum und in weichem Sinne sie mit Recht als Schlüssel anzusehen sind. Enzyme sind Elweißkörger und bestehen, wie alle Elweiße. zusammengesetzt ist, braucht uns hier nicht zu interessieren. Wichtig ist dagegen, daß von den Hunderten oder mehr verschiedenen in allen irdischen Niewesen nur 20 vorkommen. Alle Eweißarten, die es in irgendeiner auf der Erde existierenden Zeile geben mag - u gleichen 20 Aminosauren als Elementen oder Bausteinen zusammengesetzt. Passender noch als der Vergleich mit einem Baustein ist Anginanderreinung der erwähnten immer wieder gleichen 20 Aminosauren in bunter, wechseinder Aufginangerfolge. Die Jange der Ke aus Eiweißmolekulen von jeweils 51 Aminosäuren , (Auch hier sind es nur die genannten 20 verschiedenen Aminosäuren , von denen je kann. 1 Das Enzym Cythchrom c. auf dessen Bau wir näher eingehen wollen "ist ein aus 104 Aminosäuren zusammengesetztes Keth denen es besteht, ist darin durch ein bestimmtes graphisches Symbol gekennzeichnet. Konkret in dieser, und zwar genau in der hier Biochemiker Cytochronic getauft haben . Es ist , wie schon erwähnt , ein Enzym , ein * Stoffwechsel-Schlüssel * . Im Unterschied zu eine eine ganz bestimmte und nur diese eine chemische Stoffwechselreaktion auslöst (oder " aufschließt ") . in konkreten Zahlen angeben mit der die gleiche Form, die der Schlüssel hat , rein zufällig noch an ganz anderer Stelle existieren könnte . Diese Wahrscheinlichkeit ab Wahrscheinlichkeit einerfrein zufälligen "Wiederholung" entspricht hier der konkreten Change, mit der das für Cytochromic charakterist Aminosäure-Seguenz *) durch Zufal entstehen könnte "Einfacher ausgedrückt : Wie oft müßte man 104 Perlen der 20 richtigen Farbr darin rein zufällig in der Reinenfolge nebeneinandergeraten wurden "die der des Originals entspricht 7 Die Frage läßt sich präzise "mit e innerhalb einer aus 104 diedem bestehenden Kette gibt es genau 20104 verschiedene Möglichkeiten . Von diesen ist die im C reproduzieren , beträgt damit 1 zu 20104 oder , auf die gewohntere Basis 10 umgerechnet , 1 zu 10130. Damit steht fest , daß der reinen Zufall entstanden sein kann . Seit der Entstehung der Weit , seit dem "Ur-Knall" , sind erst 1017 Sekunden vergangen . (So groß vergangenen Sekunde 1 mal gewürfelt worden wäre , gäbe es heute erst höchstens 1017 verschiedene Varianten der aus 104 G ledes einzelne aller im ganzen Kosmos existierenden Moleküle eine andere Variante der lüdgliedrigen Kette repräsentierte "selbst dann geben , denn im ganzen Weltal sind nur rund 1080 Atome vorhanden! Man wird zugeben müssen , daß die Möglichkeit einer rein zufällig. beschränkten Areal der E doberfläche unter diesen im sinden mit him eichender Grieden is ausgeschlassen werden kapan Dennech aber bagegran wij (ib nam. davon weren pir is ausgegangen, auf diesen im eine mit ausgegangen, auf diesen im eine mit ausgegangen, auf diesen im eine mit ausgegangen wij (ib nam. davon weren pir is ausgegangen, auf diesen im eine mit ausgegangen, auf diesen im eine mit ausgegangen wij (ib nam. davon weren pir is ausgegangen, auf diesen im eine mit ausgegangen, auf diesen im eine mit ausgegangen wij (ib nam. davon weren pir is ausgegangen, auf diesen im eine mit ausgegangen wir (ib nam. davon weren pir ist ausgegangen, auf diesen im eine mit ausgegangen wir (ib nam. davon weren pir ist ausgegangen, auf diesen im eine mit ausgegangen wir (ib nam. davon weren pir ist ausgegangen, auf diesen im eine mit ausgegangen wir (ib nam. davon weren pir ist ausgegangen, auf diesen im eine mit ausgegangen wir (ib nam. davon weren pir ist ausgegangen wir (ib nam. davon weren pir ist ausgegangen pir ist ausgeg



nem Lahor herstellen ka diesen immer wieder noleküle - eine kettenart Insulin zum Beispiel be er Molekülkette auftreter de der Aminosäuren , au kül angeordnet , das die also mit der Oxfochrom w. Unwahrscheinlichkeit * bilden (seine * in eine Rille werfen, his riedenen Elementen durch reinen Zufall zu smos ein zweites Mal du seit Anbeginn der Welt nicht darunter . Selbst w

wide - Hendelse Ben Und Ge büher daruf untersucht worden sind i nicht nur bei uns selbst, sondem bei Und Und Ben Und Und Geraffen Und G selen . hat auch darauf noch eine Antwort . * FN * Ich habe schon erwährt . daß es sich dabe heute wirklich nur noch um eine verschwindende Minorität handelt und daß die Tatsache der Evolution heute auch von der Mehrzah derer anerkannt wird , die ihre (darwinistische) Erklärung noch immer ablehnen. "FR" Sie lautet : Es war offenbar kein Zufall , Aber ein genetischer Zusammenhang ist auch damit keineswegs bewiesen , denn Gott hat es bei eben so eingerichtet, daß alle von ihm geschaffenery alle wesen mit diesen und unzähligen anderen identischen Enzymen ausgestattet worden sind. Aber auch auf diesen Einwand läßt sich noch etwas erwidern dominated in a final charge of the second of See State Control of the Control of Unterschiede - und so geht es fort, Je entfernter die Verwandtschaft, um so größer die Zahl der Unterschiede "Genaugenommen müssten daher bei der Wahrschienlichkeitsberechnung auf Seite 48 anstatt der nur bei dem wischen weschliedenen Herscherveren gubben der 2010 lie nach der Art, mit der vergichen wird, auch niedligere Expose glein eingesetzt werden. Bei einem vergreich mit dem eines was ein 2010 3, im F des Bundes 2093 usw. Diese Korrekturen ändern aber , vor allem dann , wenn man sich nicht aleh auf den der siche bezieht , sondern die verschiedenen Arten untereinander vergleicht , an der Beweisführung grundsätzlich nichts. Die resultierenden Wahrscheinlichkeiten bleiben auch dann immer noch so extrem gering , daß die Möglichkeit , die Übereinstimmungen durch bloßen Zufäll zu erklären , weiterhin ausscheidet. Welche Ursache aber h diese individuellen Abweichungen, in denen uns dieses seiner Abstammung nach ursprünglich identische Mojekul bei den verschiedenen Arten begegnet ? Der Antwort nähern wir uns "wenn wir uns einmal vor Augen halten oft dieses Mojekul wohl reproduziert, also * kopiert * worden sein mag , seit es von der Evolution vor unausgenklich langer Zeit erfunden wurde . Bei jeder Zeiteilung mußte seine Bauanietung in allen Einzeiheiten verdoppelt worden. In jeder neuen Zeite mulitie es anhand dieser mitgegebenen Bauspian-der Bernachen vor der Bernachen der seine der nicht er nicht der nicht urworstellbaren Zeiträume, über die hinweg das von Augenblick zu Augenblick zu geschehen hat , gelingt das aber auch ihm nicht , Absolute Perfektion gibt es im Universum nicht , Auf molekularer Ebene verhindern das schol natürliche radioaktive Strahlung der Umgebung und die von der unvermeldbaren Wärmebewegung ausgehenden Störungen . Es schleichen sich also Fehler ein "Erstaunlich seiten "wenn man die Subtilität der molekularen Reproduktionsablaufe bedenkt, aber im Laufe der Zeit sammeln sie sich eben an . Da sitzt dann plötzlöt, nach der Neublidung des Molekus in einer neu entstandenen Zelle , sözusagen eine Perie *falscher Farbe * an einem bestimmten Punkt der Kette : Eine Aminosäure ist durch eine andere ausgetauscht worden . Die Bioloyen bezeichnen eine solche Anderung des erblichen Materials bekanntlich als * Mutation * . Ob der Austausch Folgen hat , I wiederum von der Stelle ab , an der die Änderung der Sequenz innerhalb der Molekülkette erfolgt ist "Getrifft die Mutation das sogenannte "aktive Zentrum" des Moleküls , dann sind die Konsequenzen - von verschwindenden Ausnahmen abgesehen - tödlich . Das aktive Zentrum ist so etwas wie der 'Bart 'des Enzym-Schlusse's , also der Teil des Mojekuls . der für seine Stoffwechsefunktion entscheidend ist. Beim Oxfochrom c besteht diese , wie schon erwährt, in der Sauerstoffübertragung innerhalb der Zeile. Eine Mutation im Bereich des jedown Zentrums von Cytochrom of ührt daher in der Regei zum sofotigen Absterben der neu entstandenen Zeile als Folge inne Erstickung : Die Mutation hat sich als 1 letal 1 erwiesen . Die verschwindende Ausnahme bedeutste in diesem Fall , daß die neue Aminosäure , die da plötzlich innerhalb dies aktiven Zentrums aufgetaucht ist , zu einer Verbesser der Funktionsfähigkeit des Moleküls führte . Daß ihr Auftreten den * Bart * des Moleküls also rein zufällig in einer Weise veränderte , welche die Sauerstoffübertragung in irgendeiner Hinsicht verbessent , indem sie sie etwa schn oder ergiebiger werden läßt. Es läßt sich leicht denken, daß das nur außergrdenden selten vorkommt. Andererseits steht ebenso fest, daß jeder einzelne derartige Fail einen Erfolg darstellt, an dem die Explution von da at hartnäckig festhalten wird. Das geschieht einfach in der Form, daß ein Disparsings, der durch einen solchen Glückstreffer in der Mutationslotterie eine Verbesserung seiner Ausstattung erfahren hat "bessere Überlebenscha hat . Fr wird daher - mit großer Wahrscheinistikeit - eine grossere Nact kommenzah hinterlassen . die . da Mutationen erbliche Anderungen sind, mit dem gleichen Vorzug ausgestattet sind. Der neue Typ setzt sich daher in e derartigen Fail auch zahlenmäßig gegenüber den weniger gut ausgestatteten Konkurrenten seiner eigenen Art innerhalb weniger Generationen durch . Kein Biologe zwelfelt heute mehr daran . daß die Spezifität eines Enzyms fe längerer Zeträume auf diesem Wege entstanden ist . Schatt für Schritt : jeder " Glückstreffer " wurde festgehalten , alle Zellen mit einer negativ wirksamen Mutation innerhalb des aktiven Zentrums (und das war ohne l die überwältigende Mehrzahl) starben dagegen ab . Sie schieden aus , af ihre Spuren sind längst getilgt . Daneben aber gab es nun noch die Mutationen , die nicht das aktive Zentrum des Moleküls betrafen , nicht den sondern sozusagen den * orff *, den bles statischen Tot des entstmatischen Molekulgerusts . Sie sind ganz offensichtlich die Ursache der Verschiedenheiten , die wir heute zwischen den Cytochrom-c-Molekulger

ell , einem insekt oder einer Filands - finden . Von den " letalen " Mutationen gibt es keine Zeugnisse mehr. Die Träger dieser Mutationen konnten nicht überleben (und ihre negab



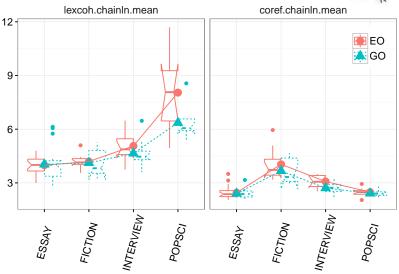
Results



Topic Development

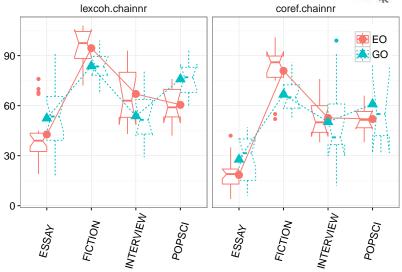
Chain Length





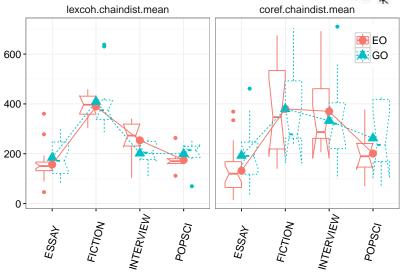
Chain Number





Chain Distance





Summary Topic Development



- language NOT significant
- language:register partially significant
- register significant
- chain types: partially different behavior

REGISTER	TOPIC DEVLOPMENT	CHAIN FEATURES
POPSCI	topic continuity within one domain	lexical chain length and distance in lexical chains and lexical chain number
ESSAY	topic variation	chain length and number and distance
FICTION	topic continuity of referents and topic interaction	coref. chain length and lexical chain lenght and chain number and chain distance
INTERVIEW	topic continuity of referents and topic variation	coref chain length and lex. chain distance



Semantic Variation

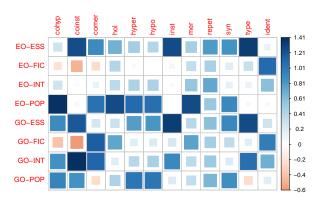
Semantic Variation and Registers



23 / 36

association between lang/registers and sem. relations (likelihood ratio)

- If ratio < 1 => log(ratio) < 0 (negative values) => red color
- If ratio > 1 = > log(ratio) > 0 (positive values) => blue color



Summary Semantic Variation



REGISTER	SEM. VARIATION	SEM. FEATURES	
POPSCI	yes! E>G	identity, (co-)instance, (co-)hyponymy, hyperonymy, holonymy, (co-)meronymy	
ESSAY	no and specific	identity, repetition, instance, type, coinstance	
FICTION	yes and specific	identity, coinstance, hyperonymy, hyponymy, GE:comeronymy	
INTERVIEW	no and specific, G>E	identity, EN:repetition, GE:type, coinstance, comeronymy	

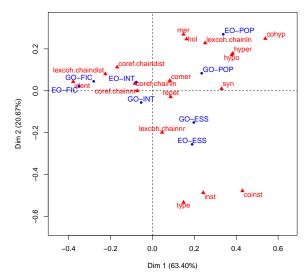


Summary

CA with All Features



CA factor map



Summary



- more differences between registers than between languages
- German registers are more alike
- English registers are more marked
- most contrasts E<=>G in POPSCI

Summary



REGISTER	TOPIC DEVLOPMENT	SEM. VARIATION
POPSCI	topic continuity within one domain	yes! E>G
ESSAY	topic variation	no and specific
FICTION	topic continuity of referents and topic interaction	yes and specific
INTERVIEW	topic continuity of referents and topic variation	no and specific, G>E



Thank you! Questions?

Information: www.gecco.uni-saarland.de

Analysed Features



- chain features (coreference, lexical)
 - chain length: number of elements in a chain
 - chain number: number of different chains
 - chain distance: between elements in the same chain
- semantic relations between adjacent chain elements
 - identity
 - repetition
 - antonymy
 - synonymy
 - hyperonymy
 - hyponymy
 - co-hyponymy

- holonymy
- meronymy
- co-meronymy
- type-instance
- instance-type
- co-instance

Annotation



Coreference: semi-automatic

- mark coreferential elements
- (distinguish antecedents and anaphors)
- (define antecedent type)
- (define anaphor type)
- assign coreferential elements to a coreference chain

Lexical chains: mostly manual

- mark lexical elements = nominal expressions
- define the type of sense relation between two adjacent chain elements
- assign lexical elements to a lexical chain



- REPETITION: orthographical repetition of nominal expressions such as *London* and *London*, or *place* and *place*.
- I live in a town called Reigate. It's between London and the countryside which is quite nice. It takes us about 25 minutes to get to London on the train. It's very nice actually, it's a nice place to live. And I grew up in a place called Banstead.
- * In case of compounding, the second element is the determining factor: stem cell but not stem cell behaviour or stem cell research.
- For instance, just identifying a true stem cell can be tricky. For scientists to be able to share results and gauge the success of techniques for controlling stem cell behavior... Of course, the goal of stem cell research is to replace...



- ANTONYMY: relation of contrast *halving* and *doubling*
- Dazu gehören zum Beispiel die Halbierung der Energie- und Rohstoffintensität bis 2020 gegenüber 1990 (bzw. 1994) und die Verdoppelung des Anteils erneuerbarer Energien am Energieverbrauch bis 2010.
- SYNONYMY: total synonymy (Lebewesen Organismen) but also near synonymy, such as between technical and common-language terms, e.g. belly and abdomen
- Wie ist die Übereinstimmung unter diesen Umständen anders zu erklären als durch eine Beziehung, die alle die Lebewesen miteinander verbindet...? Und wie anders wäre diese Beziehung zu verstehen als die einer durch Vererbung, durch "genetische Überlieferung" entstandenen Gemeinschaft von Organismen



- HYPERONYMY: in case the superordinate term follows the more specific term: village and place
- HYPONYMY: in case the specific term follows the superordinate one: place and town
- COHYPONYMY: between two elements on the same level of specification: town and village
- I live in a town called Reigate. It's between London and the countryside which is quite nice. It takes us about 25 minutes to get to London on the train. I say it's a town, it's more of a village. It's quite small. It's very nice actually, it's a nice place to live. And I grew up in a town called Banstead.



- HOLONYMY: relation, where the whole follows the part: the EU following Britain
- MERONYMY: part-whole relation, where the part follows the whole: Bulgaria following the EU
- CO-MERONYMY: succession of two parts that belong to a whole: *Romania, Croatia, Britain* following *Bulgaria*
- This means that this will not be the last enlargement of the EU. Bulgaria and Romania are waiting in the wings and likely to join in the next 2-3 years. The candidatures of Turkey and Croatia are well advanced. We in Britain welcome both. In particular we believe that Turkey, if she meets all the criteria, would add a new dimension to the EU, and represent a vital reaching out to the Islamic world at a time when such links are more than ever needed.



- TYPE: relation between a common noun and a named entity: town and Reigate
- INSTANCE: relation where the named entity follows the common noun: *Reigate* and *town*, *Banstead* and *town*
- CO-INSTANCE: relation between two named entities (Reigate and London)
- I live in a town called Reigate. It's between London and the countryside which is quite nice. It takes us about 25 minutes to get to London on the train. I say it's a town, it's more of a village. It's quite small. It's very nice actually, it's a nice place to live. And I grew up in a town called Banstead which is fairly close to Reigate.