Searching for Similarities in Text Datasets ¹

Klaus Schulz, Stefan Gerdjikov

CIS, Ludwig-Maximilians Universität München & FMI, Sofia University

26 April 2016, Göttingen

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007–2013) under REA grant agreement 625160

Table of Contents

Normalisation of Historical Texts

Extraction of Phrases

Table of Contents

Normalisation of Historical Texts

Extraction of Phrases

fol. 223v

3157

The barr of the doore, and hee this examinant sayth that hee did not see the stroke that the sayd Phelem Gaue to the said Dulin, but hee heard severall that were without the house sayinge that hee killd him with a smale skean Caled a medock, & this examinant heard the said felem Confess that hee kild him, & this examinant sawe the wownd in in the one of his brests (but which breast hee knows not) & sayth that the said Dulin spake (not that hee heard of) after the sayd wound was Giuen him, & further hee sayth that as hee thinks the wound was in or vnder the left breast & that it was not aboue an Inch brode but hee thinks it was too his hart in depth, & further sayth not Morough [mark] mc Gill Patrick his marke

This examination was taken before vs DI Huchinson Mayor Dublin

Jo: Bysse

Morrish mc Cahir Birne acknowledged t o owe vnto the keepers of the liberty of England the sume of C li. to appeare prosecution & prosecute

Figure: Instance of the 1641 Depositions' Repository of Trinity College Dublin: http://1641.tcd.ie/

fol. 223v

3157

The barr of the doore, and hee this examinant sayth that hee did not see the sirke that the sayd Pheelin Gaue to the said Dulin, but hee heard severall that were without the house sayingt that hee killd him with a smalle skean Caled a medock, at this examinant heard the said feeline Confess that hee kild him, a this examinant sawe the wown in a the one of his breats plut which breat hee knows not jk sayth that the said Dulin spake (not that hee heard of) after the sayd wound was Glieth him, a further hee sayth that as when the sayd heard wound was not written the left heard it that if was not about an inch broad but hee thinks it was too his hart in depth, & further sayth not Morough Imark inc Gall Patick his marke

The samination was laten before vs

Mayor Dubl Jo: Bysse

Morrish mc Cahir Birne acknowledged to owe unto the keepers of the liberty of England the sume of C li. to appeare prosecution & prosecute

1 Information extraction and information retrieval from historical texts is impeded due to non-standardised writing.

fol. 223v

3157

The barr of the doore, and hee this examinant saysh that hee did not see the sixties that the sayd Phelom Gaue to the said Duin, but hee heard severall that were without the house sayinge that hee killful him with a smalle skean claded a medock, did hie examinant sawe the wownin heard the said felem Confess that hee killful him with the said heek knows not? As saysh that the said following been to the breast between the heard of after the sayd wound was Guein him, A further hee sayth that as hee thinks the wound was in or winder the left breast it that it was not above an extended that the said of the saysh had say the saysh that the said of the saysh had saysh the saysh that the said the saysh had saysh the saysh that the saysh that the saysh that the saysh that saysh not Morough (hard) me of Patrick his marke

Mayor Dubl Jo: Bysse

Morrish mc Cahir Birne acknowledged to owe vnto the keepers of the liberty of England the sume of C li. to appeare prosecution & prosecute

- 1 Information extraction and information retrieval from historical texts is impeded due to non-standardised writing.
- 2 Thus, technologies relying on "correct spelling" are doomed to fail.

fol. 223v

3.10 Jam of the doore, and hee this examinant soft that hee did not see the street in the search soft the search soft that seems that were without the house sayinge that hee jild him with a smalle skean that were without the house sayinge that hee jild him with a smalle skean that were without the house sayinge that hee jild him, a this examinant sawe the women in in the one of his breats flut which breath exhause not a sayin that the said Dulm spake (not that he heard of) after the sayin wound was Cluselt him. A further hee sayth that as an inch brode but hee thinks it was too his hart in depth, & further sayin not Morough [mark] mo Gall Patick his marks

Morough [mark] mo Gall Patick his marks

Dit Huchinson

Mayor Dublin was taken before vs

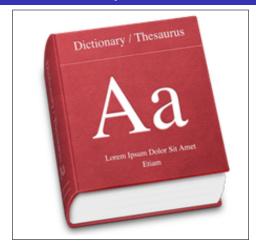
Morrish mc Cahir Birne acknowledged t o owe vnto the keepers of the liberty of England the sume of C li. to appeare prosecution & prosecute

- 1 Information extraction and information retrieval from historical texts is impeded due to non-standardised writing.
- 2 Thus, technologies relying on "correct spelling" are doomed to fail.
- 3 The spelling variations of individual words can be hardly captured in a historical lexicon.

fol. 223v 3157 The barr of the doore, and hee this examinant sayth that hee did not see the stroke that the sayd Phelem Gaue to the said Dulin, but hee heard severall that were without the house savinge that hee killd him with a smale skean Caled a medock. & this examinant heard the said felem Confess that hee kild him. & this examinant sawe the wownd in in the one of his brests (but which breast hee knows not) & sayth that the said Dulin spake (not that hee heard of) after the sayd wound was Giuen him. & further hee sayth that as hee thinks the wound was in or vnder the left breast & that it was not aboue an Inch brode but hee thinks it was too his hart in depth. & further sayth not Morough [mark] mc Gill Patrick his marke This examination was taken before vs DI Huchinson Mayor Dublin Jo: Bysse Morrish mc Cahir Birne acknowledged to owe vnto the keepers of the liberty of England the sume of C li. to appeare prosecution & prosecute

- 1 Information extraction and information retrieval from historical texts is impeded due to non-standardised writing.
- 2 Thus, technologies relying on "correct spelling" are doomed to fail.
- 3 The spelling variations of individual words can be hardly captured in a historical lexicon.

Conclusion: Historical patterns provide an implicit way to represent word variability.





List of modern words.



fol. 89r

217

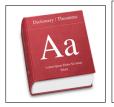
The Examynacion of Donnogh mcGillrowe of BallymaGillrowe in the Com of ffermanagh Husbandman, taken before me the 4th of 9ber 1641

<W> Whoe saith that he Came to the Cittie of Dublin friday the 22th of 8ber last, And that he Came from home the saturday before, and that Night he Lodged at a frends house of his Called Pattrick Cogglie. on the North side of Newtowne, in the foresaid Com, and there, rested vntill Munday morninge, And then he Came with some drovers of Catle the said Munday Night to a towne within 12 miles of Kells. where he lodged that Night & on <X> tuesday Night he lodged at Liscartan, and on Wednesday he went to the Navan Markett where he soulde Night he Lodged at towne within twoe miles of Navan, eight Cowes, which he had driven thither for that purpose, and on thursday night he they lodged at Killeene, and on friday Night he lodged at a house in fishamble Cookestreete, And further this Examynant saith that his Occation of Cominge to the towne was to rec. money from one, Donogh mcBryan and Owin O Brislane of Ballynacarrie in the said Com, drovers for the said 8 Cowes, which he this Examynant soulde vnto them the said drovers. And the sume which he was soe to rec. was 10 li. 5 s. but saith that he did not rec; the said money. And the Next Morninge he rose upe betymes to goe oute of the

fol. 89r 217

The Examination of Donagh mcgillrow of Ballymagillrowe in the County of fermanagh Husbandman, taken before me the 4th of November 1641

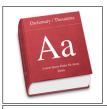
<W> Who says that he Came to the City of Dublin friday the 22th of October last, And that he Came from home the saturday before. and that Night he Lodged at a friends house of his Called Patrick Cogglie, on the North side of Newtowne, in the foresaid County. and there, rested until Monday morning, And then he Came with some drovers of Cattle the said Monday Night to a town within 12 miles of Kells, where he lodged that Night & on <X> tuesday Night he lodged at Liscartan, and on Wednesday he went to the Navan Market where he sold Night he Lodged at town within two miles of Navan, eight Cows, which he had driven thither for that purpose. and on thursday night he they lodged at Killeen, and on friday Night he lodged at a house in fishamble Cookstreet, And further this Examinant says that his Occasion of Coming to the town was to rec. money from one, Donagh mcbrian and Owen O Breslin of Ballynacarrie in the said County, drovers for the said 8 Cows. which he this Examinant sold unto them the said drovers. And the sum which he was so to rec. was 10 li. 5 s. but says that he did not rec the said money. And the Next Morning he rose up betimes to





List of modern words.

Examples of historical texts and their modern equivalents.







If hope with all my heart , and I must say this quite emphatically , that an opportunity will arise when this document can be incorporated into the Treaties at some point in the future . Should this fail to materialise , we should not be surprised if public opinion proves sceptical about Europe , or even rejects it .

The Intergovernmental Conference – to address a third subject – on the reform of the European institutions is also of decisive significance for us in Parliament .

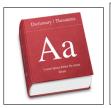
At present I feel there is a danger that if the proposal by the Belgian Government on these sanction mechanisms were to be implemented , we would be hitting first and examining only afterwards .

The Charter therefore guarantees a greater level of protection than the European Convention on Human Rights – this needs to be said to all those who are asserting the opposite here in the plenary – and it formulates political and social rights in a balanced way . We must bear in mind the Community as a whole .

But we must reach agreement on what to put in this Constitution .

I am not a Christian Democrat , so I do not believe in miracles , but I do want to salute the European Parliament as such , because it has made a major contribution to this project . In today 's debate , I associate myself with the words of all those who have spoken – Mr Moscovici , Mr Napolitano and the heads of groups – on the importance of the Charter of Fundamental Rights : this is something innovative and , moreover , unhoped-for . This does not mean that everything has to happen at once . VOTE

Report (A5-0231 / 2000) by Mr <u>Hatzidakis</u> , on behalf of the Committee on Regional Policy , Transport and Tourism , on the proposal for a European Parliament and Council directive



The Examinacion of Donnoch mcGillrows of BallymaGillrows in the Com of Remanach Mushandman, taken helves me the 4th of Sher Sper last, And that he Came from home the saturday before, and that writil Munday morninge. And then he Come with some dravers of Catle the said Munday Night to a towne within 12 miles of Kells, where he lodged that Night & on <X> tuesday Night he lodged at Liscartan, and on Wednesday he went to the Navan Markett where he soulde Night he Lodged at towns within twos miles of Navan, eight. Cowes, which he had driven thither for that purpose, and on thursday night he they locked at Killeens, and on friday Night he locked at a that his Occasion of Cominge to the towns was to rec. money from one, Donogh mcBryan and Owin O Brislane of Ballynacarrie in the said Corn, drovers for the said 8 Cowes, which he this Examunant to mc. was 10 i. 5 s. but saith that he did not rec: the said money. And the Next Mornings he rose you betyrnes to one puts of the towne, but the gates beinge shutt, he Coulde not, And scone after

was aprehended, And this Example further saith that he doth not knowe any thing of the Conspensive discovered County of Newmonth Association in Association in the Mills of Mills and Intelligence of the Mills of Mills and Intelligence of the Mills of Mills and Mills of Mills

217
The Examination of Donagh modillrow of Ballymag lirows in the

I have seed to the same of the control of the contr

The state of the control of the cont

List of modern words.

Examples of historical texts and their modern equivalents.

Modern texts

Approach (Overview), [SNG+14]



217
[The Examynacion of Donnogh mcGillrowe of BallymaGillrowe in the Com of flermanagh Husbandman, taken before me the 4th of 9ber 1641

W> Whoe suith that he Came to the Citie of Dublin friday the 22th of Sher last. And that he Came from home the saturday before, and that Night he Lodged at a frends house of his Called Pattrick Cogglie, on the North side of Newtowns in the forestid Corn and there cented yntill Munday mornings. And then he Came with some drovers of Catle the said Munday Night to a towne within 12 miles of Kells, where he lodged that Night & on <X> tuesday Night he lodged at Liscartan, and on Wednesday he went to the Navan Markett where Cowes, which he had driven thither for that purpose, and on thursday night he they lodged at Killeene, and on friday Night he lodged at a house in fishamble Cookestreete, And further this Examynant saith that his Constinu of Comings to the towns was to rec. money from said Com, drovers for the said 8 Cowes, which he this Examynant soulde vinto them the said drovers. And the sume which he was soe to rec. was 10 ii. 5 s. but saith that he did not rec: the said money. And the Next Marringe he rose upe betyrnes to goe oute of the towns, but the gates beings shutt, he Coulde not. And soons after was aprehended, And this Examynant further saith that he doth not knows any thing of the Consperacle discourred

fel. 89° 217 The Examination of Donagh mogilirow of Ballymagilirowe in the County of fermanagh Husbandman, taken before me the 4th of Novumber 1641

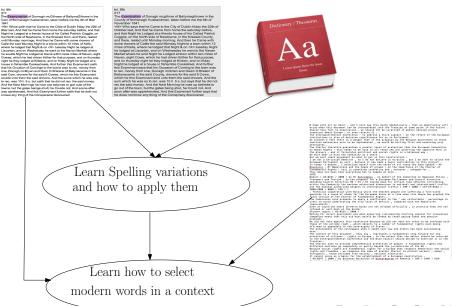
November 1641

Could be a served of the serv

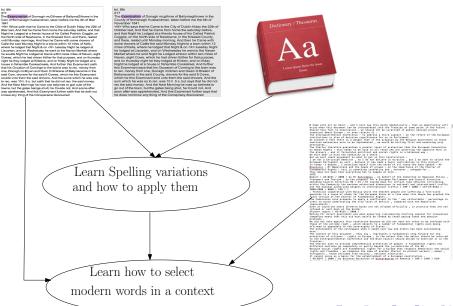


Learn Spelling variations and how to apply them

Approach (Overview), [SNG+14]



Approach (Overview), [SNG+14]



fol. 89r

The Examynacion of Donnogh mcGlirowe of BallymaGlirowe in the Corn of flarmanagh Husbandman, taken before me the 4th of 9ber

<W> Whoe saith that he Came to the Citie of Dublin friday the 22th of Sher last. And that he Came from home the saturday before, and that Night he Lodged at a frends house of his Called Patrick Cogglie, on the North side of Newtowne, in the foresaid Corn, and there, rested vntill Munday morninge. And then he Came with some drovers of Catle the said Munday Night to a towne within 12 miles of Kells. where he lodged that Night & on <>> tuesday Night he lodged at Liscartan, and on Wednesday he went to the Navan Markett where he soulde Night he Lodged at towne within twoe miles of Navan, eight Cowes, which he had driven thither for that purpose, and on thursday night he they lodged at Killeene, and on friday Night he lodged at a house in fishamble Cookestreete, And further this Examynant saith that his Occation of Cominge to the towne was to rec. money from one, Donoch mcBrvan and Owin O Brislane of Ballynacarrie in the said Corn. drovers for the said 8 Cowes, which he this Examynant. sculde virto them the said drovers. And the sume which he was soe to rec. was 10 i. 5 s. but saith that he did not rec: the said money. And the Next Morninge he rose yoe betyrnes to goe oute of the towne, but the cates beinge shuff, he Coulde not. And score after was aprehended. And this Examynent further saith that he doth not knowe any thing of the Consperacie discouered

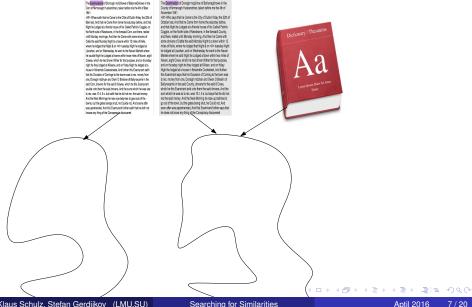
fol. 89r

The Examination of Donagh mogilirow of Ballymagilirowe in the County of fermanagh Husbandman, taken before me the 4th of

<W> Who says that he Came to the City of Dublin friday the 22th of October last. And that he Came from home the saturday before. and that Night he Lodged at a friends house of his Called Patrick Cogglie, on the North side of Newtowne, in the foresaid County, and there, rested until Monday moming, And then he Came with some drovers of Cattle the said Monday Night to a town within 12. miles of Kells, where he lodged that Night & on <X> tuesday Night he lodged at Liscartan, and on Wednesday he went to the Navan Market where he sold Night he Lodged at fown within two miles of Navan, eight Cows, which he had driven thither for that purpose, and on thursday night he they lodged at Killeen, and on friday Night he lodged at a house in fishamble Cookstreet, And further this Examinant says that his Occasion of Coming to the town was to rec. money from one. Donagh mobrian and Owen O Breslin of Ballynacame in the said County, drovers for the said 8 Cows. which he this Examinant sold unto them the said drovers. And the sum which he was so to rec. was 10 li. 5 s. but says that he did not rec the said money. And the Next Morning he rose up betimes to go out of the town, but the gates being shut, he Could not. And soon after was apprehended. And this Examinant further says that he does not know any thing of the Conspiracy discovered



61,891



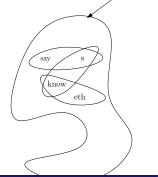
61,891

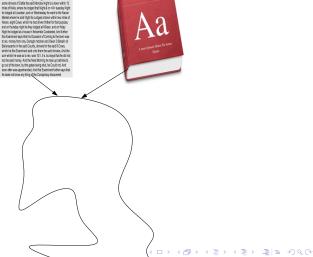


Who Whoe saith that he Came to the Citie of Dubin friday the 22th of Sher last. And that he Came from home the saturday before, and that Night he Lodged at a frends house of his Called Patrick Cogglie, on the North side of Newtowne, in the foresaid Corn, and there, rested vrtill Munday mornings. And then he Came with some dravers of Catle the said Manday Note to a towns within 17 miles of Kells where he indeed that Night & on KO buestiss Night he indeed at Liscoton, and on Wednesday he went to the Navon Markett where he soulde Night he Ladged at lowne within two miles of Navan, eight Cowes, which he had driven thither for that purpose, and on thursday night he they lodged at Killeene, and on friday Night he lodged at a house in fisherble Cookestreets. And further this Examinent saith that his Occasion of Cominge to the towne was to rec. money from one, Donogh mcBryan and Owin O Brislane of Ballynacarrie in the said Corn, dravers for the said 8 Cowes, which he this Exampnent. soulde veto them the said drovers. And the sume which he was son to rec. was 10 i. 5 s. but soft that he did not rec. the soid money. And the Next Mornings he rose you bely mos to goe outs of the towne, but the gates beinge shult, he Coulde not, And soone after was aprehended, And this Examptent further soith that he doft not knowe any thing of the Consperacial discoursed

The Examination of Donagh mogilirow of Ballymagilirowe in the County of fermanagh Husbandman, taken before me the 4th of

<W> Who says that he Came to the City of Dublin friday the 22th of October last, And that he Came from home the saturday before, and that Night he Lodged at a friends house of his Called Patrici Copplie, on the North side of Newtowne, in the foresaid County, and there, rested until Monday momino, And then he Came with some clovers of Cattle the said Monday Night to a town within 12 miles of Kells, where he ladged that Night & on <>> tuesday Night he lodged at Liscostan, and on Wednesday he went to the Navan Market where he sold Night he Lodged at lown within two miles of Navan, eight Cows, which he had driven thither for that purpose, and on thursday night he they lodged at Killeen, and on friday Night he ladged at a house in fishamble Cookstreet. And further this Examinant save that his Occasion of Coming to the town was to rec. money from one, Donagh mobrian and Owen O Breslin of Ballynacanie in the said County, drovers for the said 8 Cows, which he this Examinant sold unto them the sold drovers. And the sum which he was so to rec. was 10 li. 5 s. but says that he did not nec the said money. And the Next Morning he rose up betimes to on out of the lows, but the cates being shut, he Could not. And soon after was apprehended, And this Examinant further says that





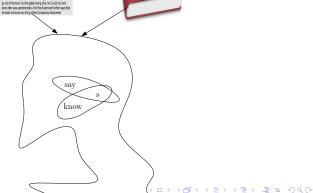
61,891



Who Whoe saith that he Came to the Citie of Dubin friday the 22th of Sher last, And that he Come from home the saturday before, and that Night he Lodged at a frends house of his Called Patrick Cogglie, on the North side of Newtowne, in the foresaid Corn, and there, rested vrtill Munday mornings. And then he Came with some dravers of Catle the said Manday Note to a towns within 17 miles of Kells where he indeed that Night & on KO buestiss Night he indeed at Liscoton, and on Wednesday he went to the Navon Markett where he soulde Night he Ladged at lowne within two miles of Navan, eight Cowes, which he had driven thither for that purpose, and on thursday night he they lodged at Killeene, and on friday Night he lodged at a house in Schemble Cookestreets. And further this Examinent saith that his Occasion of Cominge to the towne was to rec. money from one, Donogh mcBryan and Owin O Brislane of Ballynacarrie in the said Corn, dravers for the said 8 Cowes, which he this Exampnent. soulde vido them the said drovers. And the same which he was soo to rec. was 10 i. 5 s. but sails that he did not rec: the said money. And the Next Mornings he rose you bely mos to goe outs of the traine but the cates belone shuft, he Coulde not. And some other was aprehended, And this Examptent further soith that he doft not knowe any thing of the Consperacial discoursed

The Examination of Donagh mogilirow of Ballymagilirowe in the County of fermanagh Husbandman, taken before me the 4th of < W> Who says that he Came to the City of Dublin friday the 22th of October last, And that he Came from home the saturday before, and that Night he Lodged at a friends house of his Called Patrici Cogglie, on the North side of Newtowne, in the foresaid County, and there, rested until Monday momino, And then he Came with some clovers of Cattle the said Monday Night to a town within 12 miles of Kells, where he indeed that Night & on <0> tuesday Night he lodged at Liscostan, and on Wednesday he went to the Navan Market where he sold Night he Lodged at lown within two miles of Navan, eight Cows, which he had driven thither for that purpose, and on thursday night he they lodged at Killeen, and on friday Night he ladged at a house in fishamble Cookstreet. And further this Examinant save that his Occasion of Coming to the town was to rec. money from one, Donagh mobrian and Owen O Breslin of Ballynacanie in the said County, drovers for the said 8 Cows, which he this Examinant sold unto them the said drovers. And the sum which he was so to rec. was 10 li. 5 s. but says that he did not nec the said money. And the Next Morning he rose up betimes to on out of the lows, but the cates being shut, he Could not. And soon after was approhended. And this Examinant further says that





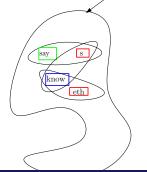
61,891

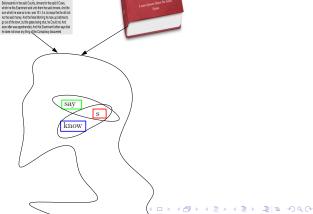


Who Whoe saith that he Came to the Citie of Dubin friday the 22th of Sher last, And that he Come from home the saturday before, and that Night he Lodged at a frends house of his Called Patrick Cogglie, on the North side of Newtowne, in the foresaid Corn, and there, rested vrtill Munday mornings. And then he Came with some dravers of Catle the said Manday Note to a towns within 17 miles of Kells where he indeed that Night & on KO buestiss Night he indeed at Liscoton, and on Wednesday he went to the Navon Markett where he soulde Night he Ladged at lowne within two miles of Navan, eight Cowes, which he had driven thither for that purpose, and on thursday night he they lodged at Killeene, and on friday Night he lodged at a house in Schemble Cookestreets. And further this Examinent saith that his Occasion of Cominge to the towne was to rec. money from one, Donogh mcBryan and Owin O Brislane of Ballynacarrie in the said Corn, dravers for the said 8 Cowes, which he this Exampnent. soulde vido them the said drovers. And the same which he was soo to rec. was 10 i. 5 s. but soft that he did not rec. the soid money. And the Next Mornings he rose you bely mos to goe outs of the towne, but the gates beinge shult, he Coulde not, And soone after was aprehended, And this Examptent further soith that he doft not knowe any thing of the Consperacial discoursed

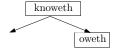
The Examination of Donagh mogilirow of Ballymagilirowe in the

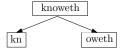
County of fermanagh Husbandman, taken before me the 4th of < W> Who says that he Came to the City of Dublin friday the 22th of October last, And that he Came from home the saturday before, and that Night he Lodged at a friends house of his Called Patrici Cogglie, on the North side of Newtowne, in the foresaid County, and there, rested until Monday momino, And then he Came with some clovers of Cattle the said Monday Night to a town within 12 miles of Kells, where he ladged that Night & on <>> tuesday Night he lodged at Liscostan, and on Wednesday he went to the Navan Market where he sold Night he Lodged at lown within two miles of Navan, eight Cows, which he had driven thither for that purpose, and on thursday night he they lodged at Killeen, and on friday Night he ladged at a house in fishamble Cookstreet. And further this Examinant save that his Occasion of Coming to the town was to rec. money from one, Donagh mobrian and Owen O Breslin of Ballynacanie in the said County, drovers for the said 8 Cows, which he this Examinant sold unto them the said drovers. And the sum which he was so to rec. was 10 li. 5 s. but says that he did not nec the said money. And the Next Morning he rose up betimes to on out of the lows, but the cates being shut, he Could not. And soon after was apprehended, And this Examinant further says that

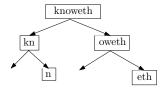


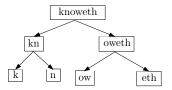


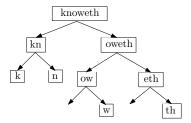
knoweth

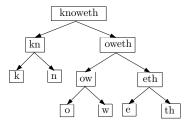


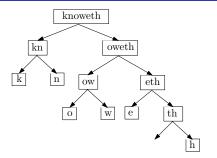


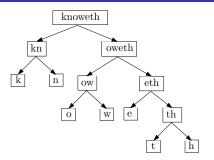


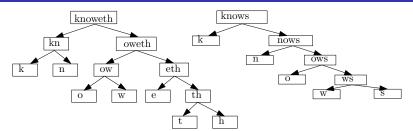


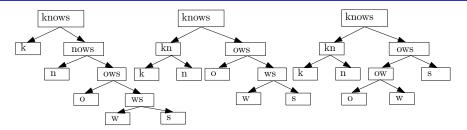


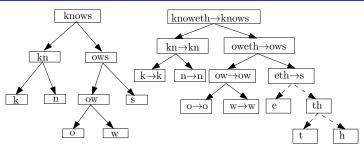






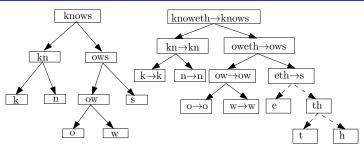






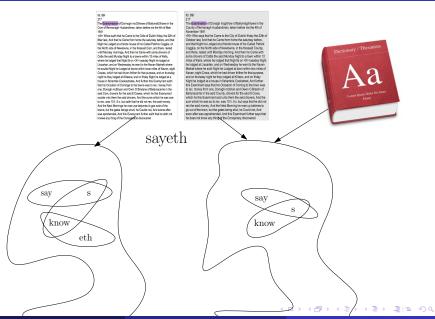
- 1 Alignment between trees.
- 2 Gathering statistics: $(\langle \eta, \nu \rangle$ historical spelling variation, $a(\eta, \nu)$ number of occurrences in all alignments)

$$\begin{array}{rcl} \mathcal{P} & = & \{ \langle \eta, \nu \rangle \, | \, \textit{a}(\eta, \nu) > 0 \} \\ \textit{p}(\nu | \eta) & = & \frac{\textit{a}(\eta, \nu)}{\sum_{\nu'} \textit{a}(\eta, \nu')} \end{array}$$



- 1 Alignment between trees.
- 2 Gathering statistics: $(\langle \eta, \nu \rangle$ historical spelling variation, $a(\eta, \nu)$ number of occurrences in all alignments)

$$\begin{array}{rcl} \mathcal{P} & = & \{ \langle \eta, \nu \rangle \, | \, \textit{a}(\eta, \nu) > 0 \} \\ \textit{p}(\nu | \eta) & = & \frac{\textit{a}(\eta, \nu)}{\sum_{\nu'} \textit{a}(\eta, \nu')} \end{array}$$



Extraction of Spelling Variations

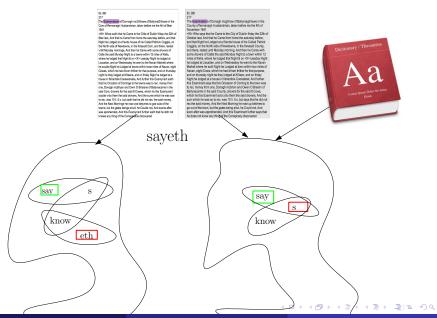


Table of Contents

Normalisation of Historical Texts

Extraction of Phrases

The amount of electronically available plain texts boosts.

The amount of electronically available plain texts boosts.

The access to the information in these texts is hindered by the lack of explicit structure.

The amount of electronically available plain texts boosts.

The access to the information in these texts is hindered by the lack of explicit structure.

Some of these texts are similar but finding this out is costly and time consuming.

The amount of electronically available plain texts boosts.

The access to the information in these texts is hindered by the lack of explicit structure.

Some of these texts are similar but finding this out is costly and time consuming.

Idea: Use the orthographic similarities among these texts to induce structure and facilitate search queries.

Required Resources

In the presence of cycloheximide two types of RNA accumulate in the cell: soluble RNA and a high-molecular-weight RNA. The latter has a base composition intermediate between those of yeast DNA and yeast ribosomal RNA, and sediments in a suc gradient at a rate faster than that of the 23s ribosomal RNA component.

4. Yeast ribosomal RNA contains methylated bases.

Judged from the incorporation of [Me-(14)C]methionine, the extent of methylation of ribosomal RNA is about 20% of that of

Judged from the incorporation of [Me-(14)C]methionine, the extent of methylation of ribosomal RNA is about 20% of that of the ;soluble' RNA fraction.

The high-molecular-weight RNA formed in the presence of cycloheximide is less methylated than normal RNA.

In this case the sucrose-density-gradient sedimentation patterns of newly methylated and newly <u>synthesized</u> RNA do not coi In the presence of cycloheximide, polysomal material accumulates, indicating that messenger RNA is formed.

The effect of the antibiotic on protein and RNA synthesis can be abolished by washing of the cells.

The RNA that has accumulated during incubation of the cells with the antibiotic is not stable on removal of cycloheximide

The results presented in this study are discussed in relation to the regulation of RNA formation in yeast.

1. Methods are described for measuring tidal volume and frequency, end-tidal CO(2), blood pressure and heart rate, and ar

tensions in the <u>unanaesthetized</u> new-born lamb.2.

The resting values of minute ventilation (V)/kg body wt. and arterial oxygen and carbon dioxide tension, (P(a, o2)) and (V)

were similar to those which have been reported in the new-born baby.

There was a direct and significant relation between P(a, o2) and P(a, CO2) and the age of the lamb.3.

Thirty-five unanaesthetized lambs aged 40 min to 10 days breathed 100% oxygen; minute ventilation fell by an average of 1 control, end-tidal CO(2) increased and the ratio of change in tidal volume (DeltaP) (To change in pressure (DeltaP) (De DeltaP) remained constant.

The effect of breathing 100% O(2) on lung compliance was variable.4.

These changes in ventilation were virtually abolished after both sinus nerves had been cut.5.

The results therefore suggest that a significant hypoxic drive to ventilation exists in the new-born lamb and that this d mediated by functioning and mature peripheral chemoreceptors.6.

Preliminary evidence suggested that, on 100% O(2), the sensitivity of new-born lambs to inhaled CO(2) was reduced.

The molecular weight and amino acid composition of triose phosphate isomerase have been determined.

The molecular weight (43000) is lower and the molecular activity (500000) higher than those of most other glycolytic enzy. Reaction with <a href="mailto:index.org/linear.com/li

obtained are consistent with historie's being the basic group in the active centre.

1. <u>LMandelate</u> dehydrogenase and <u>mandelate racemase</u> were partially purified from extracts of Pseudomonas <u>fluorescens</u> A-31 media containing d-mandelate.

media containing <u>d-mangelare</u>.

The activity of <u>mangelare racemase</u>, but not that of <u>l-mandelare</u> dehydrogenase, is greatly stimulated by Mg(2+), Mn(2+), C though less effectively. by Ni(2+).

Other metal ions are inactive or inhibitory.

Racemase activity is inhibited by phosphate, fluoride, pyrophosphate and EDTA.

Required Resources

```
In the presence of cycloheximide two types of RNA accumulate in the cell: soluble RNA and a high-molecular-weight RNA.
The latter has a base composition intermediate between those of yeast DNA and yeast ribosomal RNA, and sediments in a sucrose
gradient at a rate faster than that of the 23s ribosomal RNA component.
4. Yeast ribosomal RNA contains methylated bases.
Judged from the incorporation of [Me-(14)C]methionine, the extent of methylation of ribosomal RNA is about 20% of that of
the :soluble' RNA fraction.
The high-molecular-weight RNA formed in the presence of cycloheximide is less methylated than normal RNA.
In this case the sucrose-density-gradient sedimentation patterns of newly methylated and newly synthesized RNA do not coincide.
In the presence of cycloheximide, polysomal material accumulates, indicating that messenger RNA is formed.
The effect of the antibiotic on protein and RNA synthesis can be abolished by washing of the cells.
The RNA that has accumulated during incubation of the cells with the antibiotic is not stable on removal of cycloheximide.
The results presented in this study are discussed in relation to the regulation of RNA formation in yeast.
1. Methods are described for measuring tidal volume and frequency, end-tidal CO(2), blood pressure and heart rate, and arterial gas
tensions in the unanaesthetized new-born lamb.2.
The resting values of minute ventilation (V)/kg body wt. and arterial oxygen and carbon dioxide tension, (P(a, o2)) and (P(a, CO2))
were similar to those which have been reported in the new-born baby.
There was a direct and significant relation between P(a, o2) and P(a, CO2) and the age of the lamb.3.
Thirty-five unanaesthetized lambs aged 40 min to 10 days breathed 100% oxygen; minute ventilation fell by an average of 19% of
control, end-tidal CO(2) increased and the ratio of change in tidal volume (DeltaV(T)) to change in pressure (DeltaP) (DeltaV(T)/
DeltaP) remained constant.
The effect of breathing 100% O(2) on lung compliance was variable.4.
These changes in ventilation were virtually abolished after both sinus nerves had been cut.5.
The results therefore suggest that a significant hypoxic drive to ventilation exists in the new-born lamb and that this drive is
mediated by functioning and mature peripheral chemoreceptors.6.
Preliminary evidence suggested that, on 100% O(2), the sensitivity of new-born lambs to inhaled CO(2) was reduced.
The molecular weight and amino acid composition of triose phosphate isomerase have been determined.
The molecular weight (43000) is lower and the molecular activity (500000) higher than those of most other glycolytic enzymes.
Reaction with iodoacetate (studied with radioactive reagent) takes place in two phases: in the first phase, at pH6.3, cysteine and
methionine groups react and enzymic activity is unimpaired; in the second phase, histidine reacts and enzymic activity is lost.
Photo-oxidation leads to inactivation, with loss of cysteine, of histidine and of tryptophan, but little loss of tyrosine.
The mechanism postulated for the action of the enzyme demands the intervention of a group functioning as a base, and the results
obtained are consistent with histidine's being the basic group in the active centre.
1. I-Mandelate dehydrogenase and mandelate racemase were partially purified from extracts of Pseudomonas fluorescens A-312 grown in
media containing d-mandelate.
The activity of mandelate racemase, but not that of 1-mandelate dehydrogenase, is greatly stimulated by Mg(2+), Mg(2+), Co(2+) and,
though less effectively, by Ni(2+).
Other metal ions are inactive or inhibitory.
```

Set of sentences or paragraphs or other meaningful units of texts.

3. Racemase activity is inhibited by phosphate, fluoride, pyrophosphate and EDTA.

Young children recognise function words but do not use them in the beginning. [HW03, SWC06, Tra07]

(A1) Phrases occur in different contexts.

- (A1) Phrases occur in different contexts.
- (A2) Sentences are built of (overlapping) phrases.

- (A1) Phrases occur in different contexts.
- (A2) Sentences are built of (overlapping) phrases.
- (A3) Phrases overlap on function words.

- (A1) Phrases occur in different contexts.
- (A2) Sentences are built of (overlapping) phrases.
- (A3) Phrases overlap on function words.
- (P1) The function words occur regularly as the boundaries of phrases in the corpus.

- (A1) Phrases occur in different contexts.
- (A2) Sentences are built of (overlapping) phrases.
- (A3) Phrases overlap on function words.
- (P1) The function words occur regularly as the boundaries of phrases in the corpus.
- (P2) A particular sentence should exhibit these regularities as good as its structure w.r.t. the corpus allows.

Approach (2)

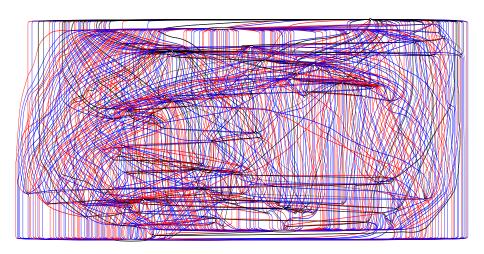
In the presence of cycloheximide, polysomal material accumulates, indicating that messenger RNA is formed.

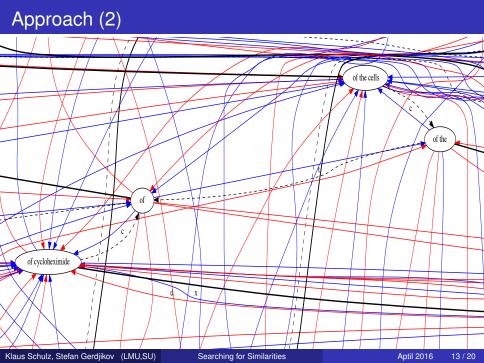
The effect of the antibiotic on protein and RNA synthesis can be abolished by washing of the cells.

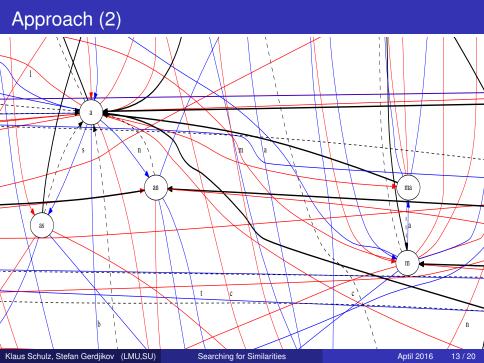
The RNA that has accumulated during incubation of the cells with the antibiotic is not stable on removal of cycloheximide.

Excerpt from the Medline Corpus: http://mbr.nlm.nih.gov/

Approach (2)







Approach (2)

$$\begin{array}{lll} \textit{pref}(F|\mathcal{P}) & = & |\{F \circ S \in \mathcal{P} \mid S \in \Sigma^+\}| \\ \textit{suff}(F|\mathcal{P}) & = & |\{S \circ F \in \mathcal{P} \mid S \in \Sigma^+\}| \\ \textit{inf}(F|\mathcal{P}) & = & |\{S_1FS_2 \in \mathcal{P} \setminus \{F\} \mid (S_1,S_2) \in (\Sigma^*)^2\}| \\ p_{\textit{pref}}(F|\mathcal{P}) & = & \frac{\textit{pref}(F|\mathcal{P})}{\textit{inf}(F|\mathcal{P})}, p_{\textit{suf}}(F|\mathcal{P}) = \frac{\textit{suff}(F|\mathcal{P})}{\textit{inf}(F|\mathcal{P})} \end{array}$$

```
 \begin{array}{ll} \text{reflect (P1):} & p_{\mathit{fw}}(F|\mathcal{P}) = p_{\mathit{pref}}(F|\mathcal{P}) \cdot p_{\mathit{suf}}(F|\mathcal{P}) \\ \text{reflect (P2):} & \ell(\overline{F}|S_n,\mathcal{P},\mathcal{C}) = \prod_{i=1}^{k_n-1} p_{\mathit{fw}}(F_i|\mathcal{P}) \\ \text{refine phrases:} & \mathcal{P}_0 = \mathcal{G}(\mathcal{C}) & \mathcal{P}_{i+1} = \mathcal{P}'_i & \text{until } \mathcal{P}_i \sim \mathcal{P}_{i+1} \\ \end{array}
```

 $\mathcal{P}' := \{S_n[j_1..j_2] \mid S_n[j_1..j_2] \text{ optimal phrase candidate for a sentence } S_n\}$

Results

Corpus	# of funct. words	most frequently used function words		
WikiEN	305	_; the ;. ; and ; of ; in ; to ; a ; of the ; is ;" ; was ;. The ; in the ;		
		for; that; by; as; with; are;-;); on; were; (; to the; from; at; In;		
EU-Parl	200	_; the ; , ; to ; of ; a ; and ; in ; that ; be ; for ; this ; which ; in the ;		
(EN) 1%		are; we; will; have; on; has; an; not; to the; with; , the;		
Medline	146	that the; it; as; by; to be;, and; more; _; of; and; the; in; of the; to; was; with; for; a; were; is; in the;		
(10%)	140	by : that : are :), : from : on : at : or : after : as : +/- : and the :		
(1070)		between; during; to the;;; patients with;		
Medline	143	_; of ; the ; and ; in ; to ; of the ; was ; a ; were ; with ; for ; is ; in the ;		
(0.5%)		by; that; are; from;),; at; on; as; after; to the; or; an;		
		patients; +/-; cells;;; and the;		
WikiDE	305	_;. ; der ; die ; und ; des ; in ; von ; "; -; im ; das ;, ;) ; eine ; den ;		
		als;. Die; zu; mit; ein; ist; in der; wurde; für; auf; zur; einer;		
EU-Parl 1%	159	nach; sind;. Der; nicht;, der; _;,; die; der; und; zu; in; den; eine;, daß; des;, die; das;		
(DE)	109	nicht; von; werden; im; ist; ein; für; wir; auf; es; einer; auch;		
(52)		zur; mit; dem; sich;		
Wittgenstein	200	_;. ;[; die ; der ; das ;[1];? ; the ; & ; daß; in ; nicht ; eine ; ein ; ist ;		
(DĚ+EN)		den; to; es;,; zu;;; a; so; von; of; that; wenn; was; I;		
		er; als; einen; wie; aber; sie;		
Accountant	379	_; de; een; en.;.; en; in; dat; is;,; die;. De; (; niet; op;		
(NL)	454	;; in de; van het; zijn; voor;); ^; met; als; dat de;. Het; wordt;		
EU-Parl (NL) 1%	151	_; , ; van ; een ; het ; in ; en ; van de ; dat ; te ; op ; die ; is ; niet ; voor ; zijn ; de ; worden ; met ; om ; dat de ; deze ; heeft ; van het ;		
(INL) 1/0		aan; voor de; ook; er; wij;		
EU-Parl	203	_;, ; de ; des ; et ; les ; la ; à ; que ; en ; pour ; une ; du ; qui ; dans ;		
(FR) 100%				
` ,		ont; sur; d'une;, les; plus; d'un;		
EU-Parl	173	_;, ;que; la; en; y; los; el; a; para; del; una; se; un; las; con;		
(ES) 100%		no; al; a la; es; por; en la; de las; a los;, que;, y; ha; más; su;		

Results

The RNA that has accumulated during incubation of the cells with

the antibiotic is not stable on removal of cycloheximide.

Results

The RNA that has accumulated during incubation of the cells with

has accumulated during

The RNA that

incubation of the cells with the

the antibiotic is not stable on removal of cycloheximide.

the antibiotic is

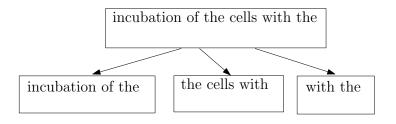
on removal of

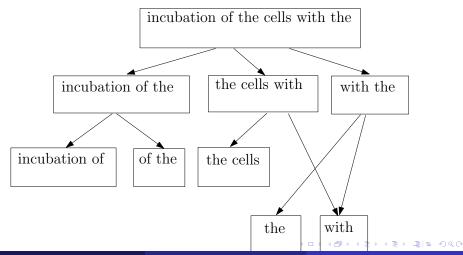
is not stable

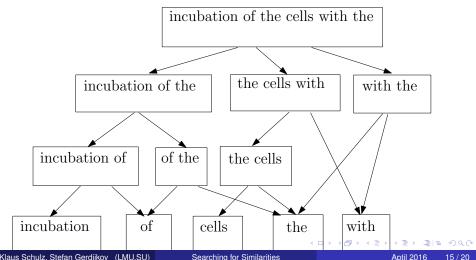
of cycloheximide.

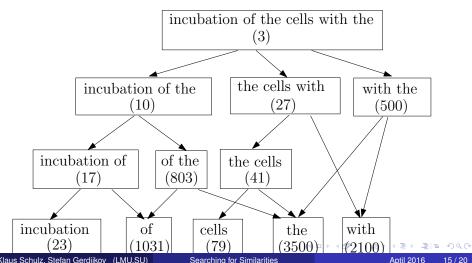
Hierarchical decomposition of phrases.

incubation of the cells with the

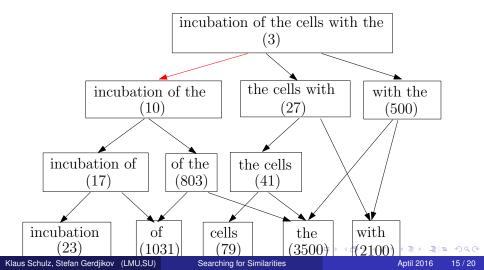




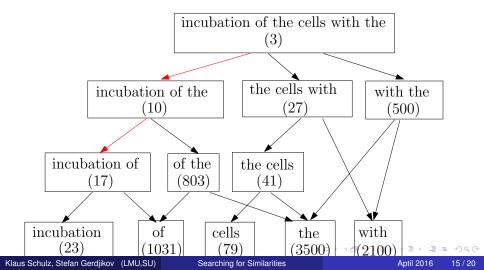




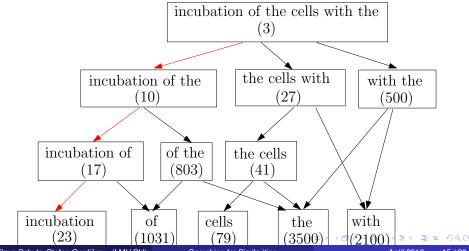
Hierarchical decomposition of phrases.



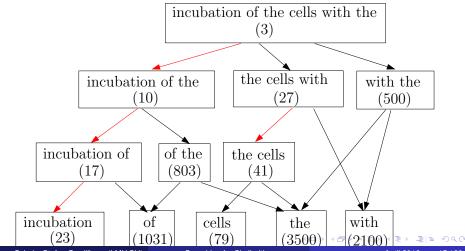
Hierarchical decomposition of phrases.



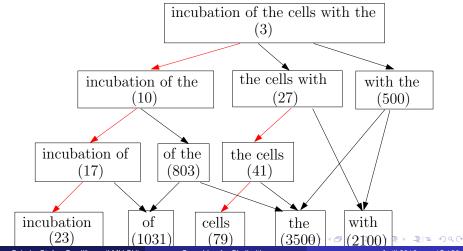
Hierarchical decomposition of phrases.



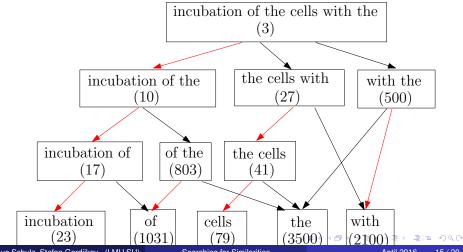
Hierarchical decomposition of phrases.



Hierarchical decomposition of phrases.



Hierarchical decomposition of phrases.



Hierarchical decomposition of phrases.

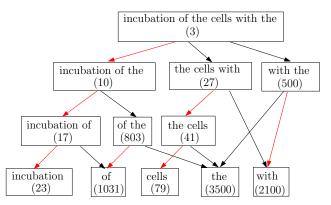
 observed 	21. one	41. tumor
2. those	22. similar	42. children
3. used	23. performed	43. identified
4. only	24. control	44. use
treatment	25. lower	45. different
6. groups	26. demonstrated	46. examined
7. present	27. disease	47. normal
8. obtained	28. age	48. time
9. increase	29. cells	49. blood
increased	30. data	50. cases
reported	31. reduced	51. studied
12. found	32. decreased	52. specific
results	33. risk	53. subjects
14. higher	34. changes	54. decrease
patients	35. group	55. protein
significantly	36. previously	56. 3
17. patient	developed	57. compared
18. cancer	38. high	58. women
19. study	39. effective	59. investigated
20. expression	40. factors	60. evaluated

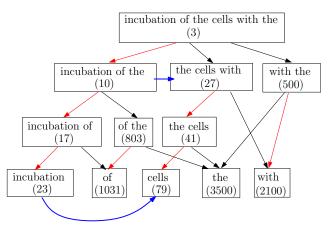
Hierarchical decomposition of phrases.

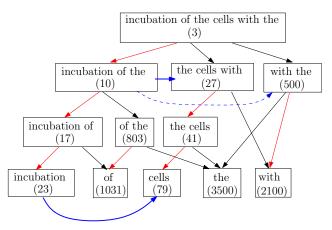
- 1. observed (2310/602) 2. those (121/30) 3. used (220/1) 4. only (9/5) 5. treatment (2223/1106) 6. groups (361/142) 7. present (127/114) 8. obtained (3743/1141) 9. increase (534/358) 10. increased (991/386) 11. reported (4701/340) 12. found (719/8) 13. results (723/432) 14. higher (1253/505) 15. patients (4104/1897) 16. significantly (6063/958) 17. patient (-/3414) 18. cancer (-/2935) 19. study (1271/134) 20. expression (1333/1666)
- 21. one (17/3) 22. similar (1341/213) 23. performed (5841/192) 24. control (362/150) 25. lower (3088/550) 26. demonstrated (1167/1900) 27. disease (3185/1674) 28. age (1224/896) 29. cells (-/1123) 30. data (1947/277) 31. reduced (1340/773) 32. decreased (5109/3881) 33. risk (556/1182) 34. changes (728/270) 35. group (198/66) 36. previously (3324/676) 37. developed (633/26) 38. high (532/226) 39. effective (416/522) 40. factors (1920/1089)
- 41. tumor (-/10737) 42. children (1001/177) 43. identified (1864/608) 44. use (151/6) 45. different (231/194) 46. examined (4705/5826) 47. normal (3079/1211) 48. time (41/2) 49. blood (-/1376) 50. cases (561/316) 51. studied (4507/846) 52. specific (284/462) 53. subjects (1068/2110) 54. decrease (3735/2992) 55. protein (7516/1941) 56. 3 (720/779) 57. compared (2692/1199) 58. women (276/199) 59. investigated (3658/5008)

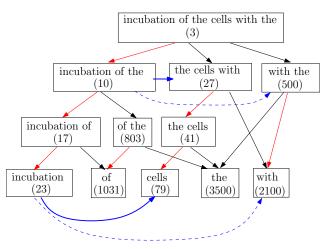
Word Fields (1)

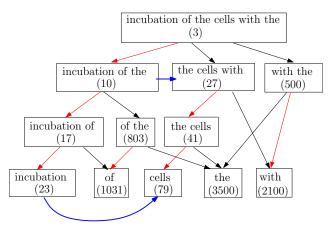
The hierarchical decomposition of phrases allows us to relate cooccurring atomic phrases.



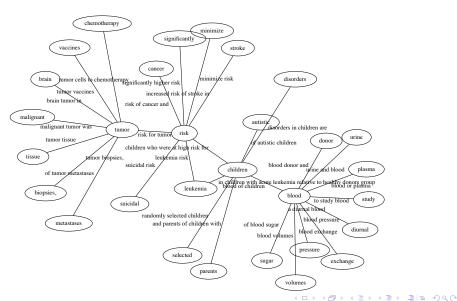




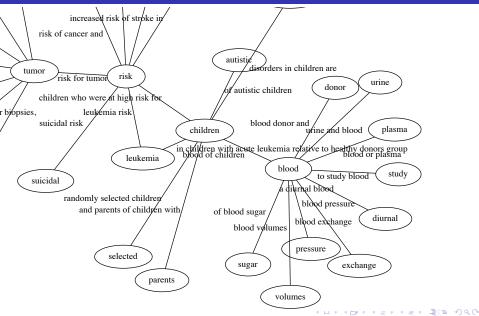




Word Fields (Medline Corpus)

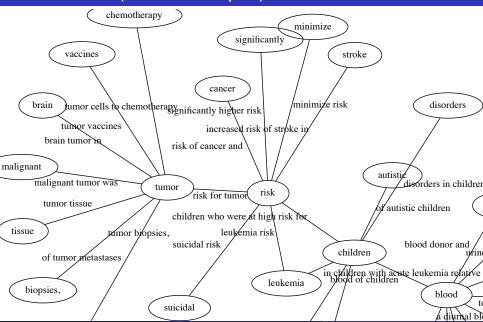


Word Fields (Medline Corpus)



Word Fields (Medline Corpus)

Klaus Schulz, Stefan Gerdiikov (LMU.SU)

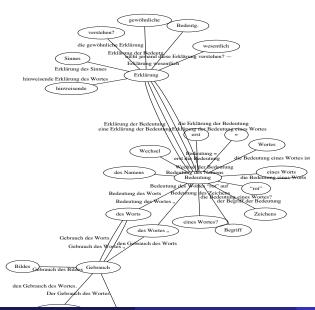


Searching for Similarities

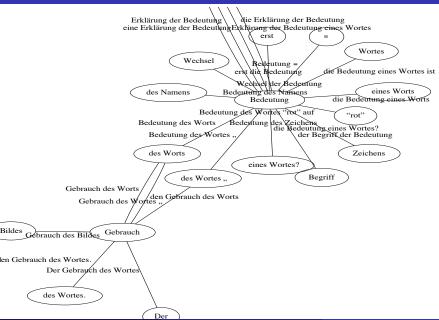
Aptil 2016

17/20

Word Fields (Wittgenstein Corpus)



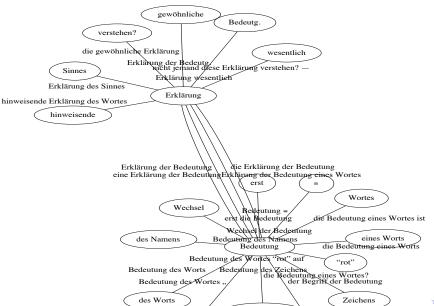
Word Fields (Wittgenstein Corpus)



Bildes

Aptil 2016

Word Fields (Wittgenstein Corpus)





S. Gerdjikov, S. Mihov, and V. Nenchev.

Extraction of spelling variations from language structure for noisy text correction.

In ICDAR'13, pages 324-328, 2013.



S. Gerdjikov and K. U. Schulz.

Corpus analysis without prior linguistic knowledge - unsupervised mining of phrases and subphrase structure.

CoRR, abs/1602.05772, 2016.



Barbara Höhle and Jürgen Weissenborn.

German-learning infants' ability to detect unstressed closed-class elements in continuous speech.

Developmental Science, pages 122-127, 2003.



A. Sariev, V. Nenchev, S. Gerdjikov, P. Mitankin, H. Ganchev, S. Mihov, and T. Tinchev.

Flexible noisy text correction.

In DAS'14, pages 31-35. IEEE, 2014.



Rushen Shi, Janet Werker, and Anne Cutler.

Recognition and representation of function words in english learning infants.

Infancy, 10(2):187-198, 2006.



Rosemarie Tracy.

Wie Kinder Sprachen Lernen: Und wie wir sie dabei unterstützen können.

Nar Francke Attempto Verlag, 2007.

Corpus	$ \mathcal{F}(\mathcal{C}) $	most frequently used function words
保镖天下	293	; ; , ; 的 (of); 了 (over); 是 (is/are); 你(you); 我(I/me); 这(this); 人(people/human); 着(_); 在(at/in);
Chinese		就(_); ""; 黎箫(Li Xiao – a person's name in this fiction); 他(he); 也(also/too); 到(arrive); ? (?);
Fiction		都(both/all); 去(go);! (l); 来(come); 还(still/again); 看着(look/stare); 那(that); 一个(one);
		出(go/get out); 上(up/above); 要(need); 和(and); 自己(myself); 将(will/shall); 没有(no/didn't);

Corpus	$ \mathcal{F}(\mathcal{C}) $	most frequently used function words
保镖天下	293	; ; , ; 的 (of); 了 (over); 是 (is/are); 你(you); 我(I/me); 这(this); 人(people/human); 着(_); 在(at/in);
Chinese		就(_); ""; 黎箫(Li Xiao – a person's name in this fiction); 他(he); 也(also/too); 到(arrive); ? (?);
Fiction		都(both/all); 去(go);! (l); 来(come); 还(still/again); 看着(look/stare); 那(that); 一个(one);
		出(go/get out); 上(up/above); 要(need); 和(and); 自己(myself); 将(will/shall); 没有(no/didn't);

Corpus	$ \mathcal{F}(\mathcal{C}) $	most frequently used function words
保镖天下	293	; ; , ; 的 (of); 了 (over); 是 (is/are); 你(you); 我(I/me); 这(this); 人(people/human); 着(_); 在(at/in);
Chinese		就(_); ""; 黎箫(Li Xiao – a person's name in this fiction); 他(he); 也(also/too); 到(arrive); ? (?);
Fiction		都(both/all); 去(go);! (l); 来(come); 还(still/again); 看着(look/stare); 那(that); 一个(one);
		出(go/get out); 上(up/above); 要(need); 和(and); 自己(myself); 将(will/shall); 没有(no/didn't);

Corpus	$ \mathcal{F}(\mathcal{C}) $	most frequently used function words
保镖天下	293	; ; , ; 的 (of); 了 (over); 是 (is/are); 你(you); 我(I/me); 这(this); 人(people/human); 着(_); 在(at/in);
Chinese		就(_); ""; 黎箫(Li Xiao – a person's name in this fiction); 他(he); 也(also/too); 到(arrive); ? (?);
Fiction		都(both/all); 去(go);! (l); 来(come); 还(still/again); 看着(look/stare); 那(that); 一个(one);
		出(go/get out); 上(up/above); 要(need); 和(and); 自己(myself); 将(will/shall); 没有(no/didn't);

Corpus	$ \mathcal{F}(\mathcal{C}) $	most frequently used function words
保镖天下	293	; ; , ; 的 (of); 了 (over); 是 (is/are); 你(you); 我(I/me); 这(this); 人(people/human); 着(_); 在(at/in);
Chinese		就(_); ""; 黎箫(Li Xiao – a person's name in this fiction); 他(he); 也(also/too); 到(arrive); ? (?);
Fiction		都(both/all); 去(go);! (l); 来(come); 还(still/again); 看着(look/stare); 那(that); 一个(one);
		出(go/get out); 上(up/above); 要(need); 和(and); 自己(myself); 将(will/shall); 没有(no/didn't);