Nuevas herramientas en la lingüística diacrónica: tradiciones discursivas y lingüística computacional

Giulia Mazzola & Stefano De Pascale

KU LEUVEN

Proyecto C14/18/034 'Beyond the clause' financiado por el Research Council de la KU Leuven

Contenido

- Tradiciones discursivas y lingüística de corpus cuantitativa
- Modelos de espacio vectorial
- Detección automática de grupos documentales en el CODEA
- Aplicaciones concretas
- Conclusiones

Preguntas de investigación

- ¿Cómo se puede compaginar el modelo de las TDs con el uso de los grandes corpus? ¿En qué criterios uno se puede basar para definir e identificar una TD?
- ¿Siempre es necesario diferenciar entre las TDs y otros conceptos relacionados con la variación textual (por ejemplo, géneros, tipos textuales y registros)?

El corpus

- **CODEA** ofrece en su estado actual 2500 documentos en español de toda la geografía peninsular del español, anteriores al año 1800
- Incluye un espectro social amplio, que va del documento regio a la nota humilde, incluyendo las llamadas “manos inhábiles”
- Un corpus fiable desde el punto de vista de la transcripción filológica
- Alcance temporal para este estudio: 1400-1799

Variables de clasificación documental incluidas en el corpus (CHARTA)

- Ámbito textuales (Cancilleresco, Eclesiástico, Judicial, Municipal, Particular)
- Tipos documentales (Actas, Certificaciones, Cartas privadas, Textos legislativos, Estatutos, Testamentos, Informes, Notas, Compraventa, Contratos…)

Conceptos llaves

- La lengua no es un objeto monolítico (Kabatek et al. 2010, Kabatek 2005, 2013)

- Cada texto permite multiples maneras de clasificación global o interna, y todas corresponden a TDs (Kabatek 2013:19)

- La frecuencia relativa de elementos gramaticales, lexicales y formularios puede ser una indicación de TDs diferentes (cf. Kabatek et al 2010:7)

Propuesta

Acercamientos estadísticos y computacionales

- Biber (1988) y Biber et al. (2006)
- Kabatek et al. (2010)
- Mazzola et al. (2020)

Clasificación textual desde abajo hacia arriba

→ utilizando las frecuencias y la distribución de las palabras en cada texto

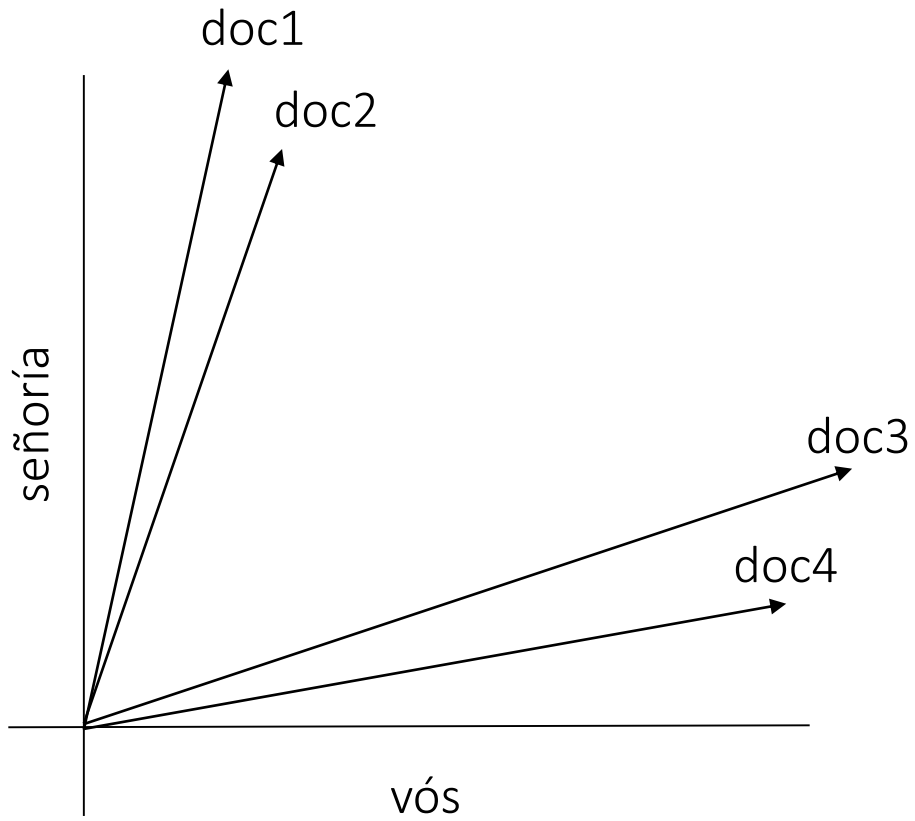
→ comparando las semejanzas de estos perfiles de frecuencia para agrupar textos.

Modelo de espacio vectorial

Cf. De Pascale (2019)

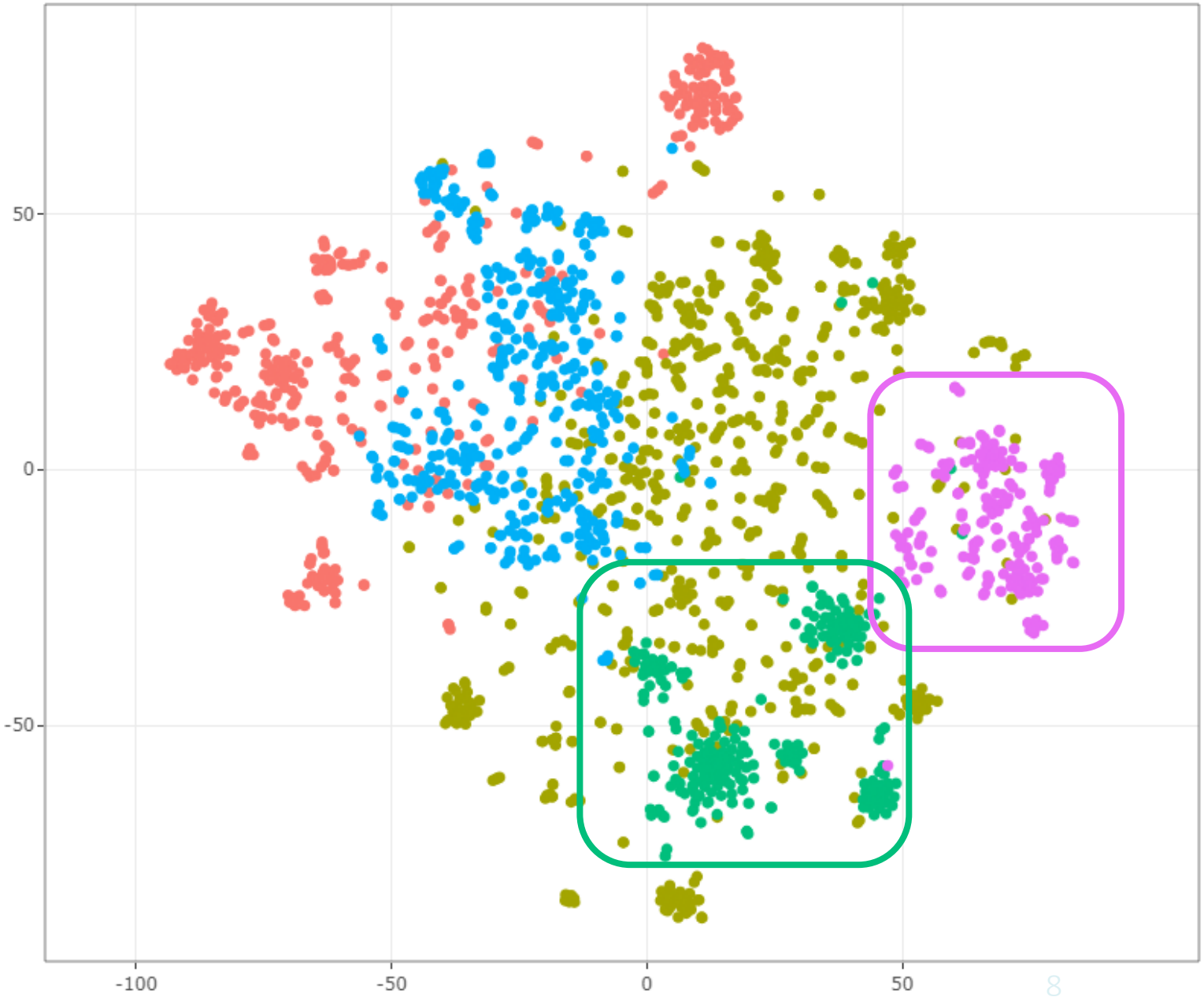
	señoría	merced	señor	convento	vós	ciudad
doc1	284	79	125	134	0	3
doc2	163	178	114	52	0	0
doc3	8	4	5	7	123	184
doc4	0	2	0	1	360	354

Vector del documento



Modelo de espacio vectorial

- 1 punto = 1 documento
- puntos cercanos = textos parecidos
- Agrupamiento jerárquico basado en estas similitudes
- 5 grupos



Modelo de espacio vectorial

Afinación de los criterios

Criterio interno:

- **Unigramas** (= palabras sueltas, p.e.: *señoría, merced, convento, vós, ciudad*…)

- **Bigramas/Trigramas** (=dos/tres palabras contiguas, p.e.: *notario público, siempre jamás, vuestra señoría ilustrísimas, estos nuestros reinos*…)

Modelo de espacio vectorial

Skipgrams (=palabras contiguas con huecos) → colocaciones con huecos variables

- **Está___ y se llama ___**

Está **cristiano** y se llama **Lorenço** (CODEA-1650)

Esta niña está **cristiana** y se llama **Francisca Gerónima**. (CODEA-1568)

- **____los pies _____ de vuestra**

Besamos los pies de vuestra magestad (CODEA-1102)

Beso los pies y manos de vuestra magestad (CODEA-1265)

Modelo de espacio vectorial

Afinación de los criterios

Frecuencias

- Remoción de las palabras demasiado frecuentes (100%-70% de los documentos; las 50 palabras más frecuentes en absoluto) → **son poco informativas** (p.e.: de, e, que, y, la, en, a…)
- Remoción de las palabras demasiado infrecuentes (<10 ocurrencias, p.e.: dental, oliva, líquidos….)

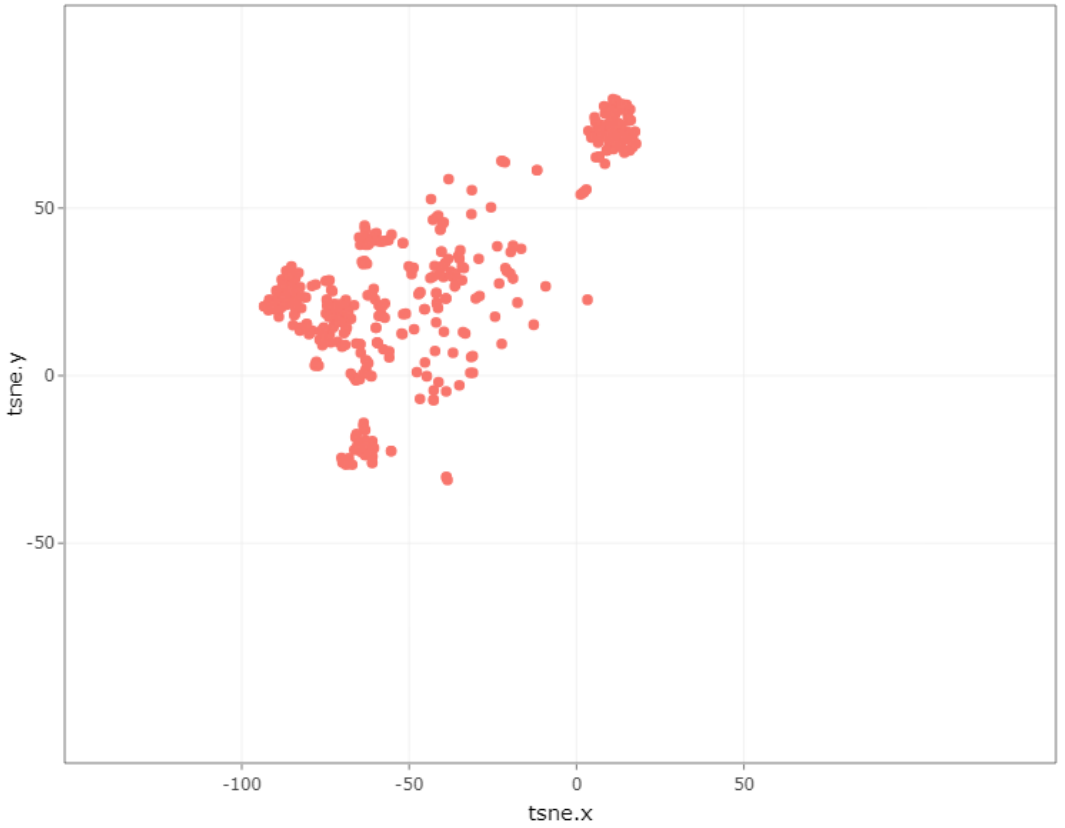
Unigramas

- Criterio interno: palabras sueltas
- Frecuencia mínima= 3 ocurrencias
- Frecuencia máxima= exclusión de palabras que aparecen en más del 70% de los documentos

Unigramas

Grupo	Etiqueta	N. Documentos
1	Latín y Oriental	48

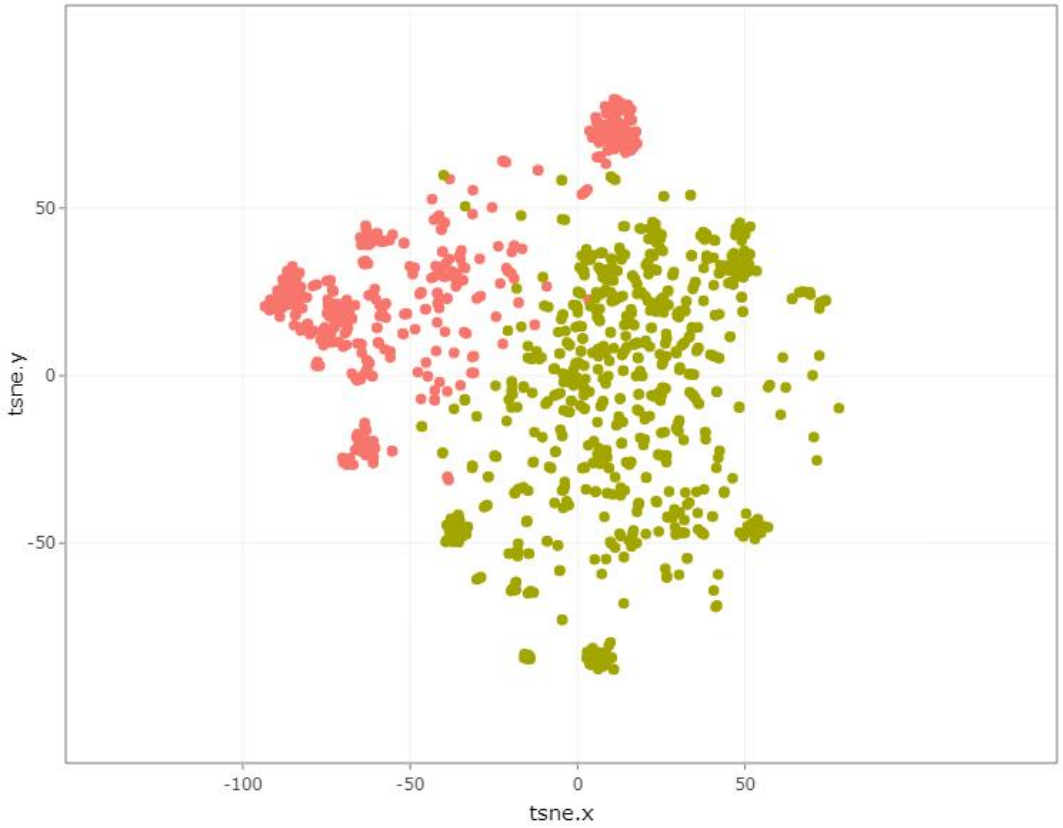
Palabras típicas: *et, testist, confirmat, cum, ciudat, testament, dito, don*



Unigramas

Grupo	Etiqueta	N. Documentos
1	Latín y Oriental	48
2	Burocrático	508

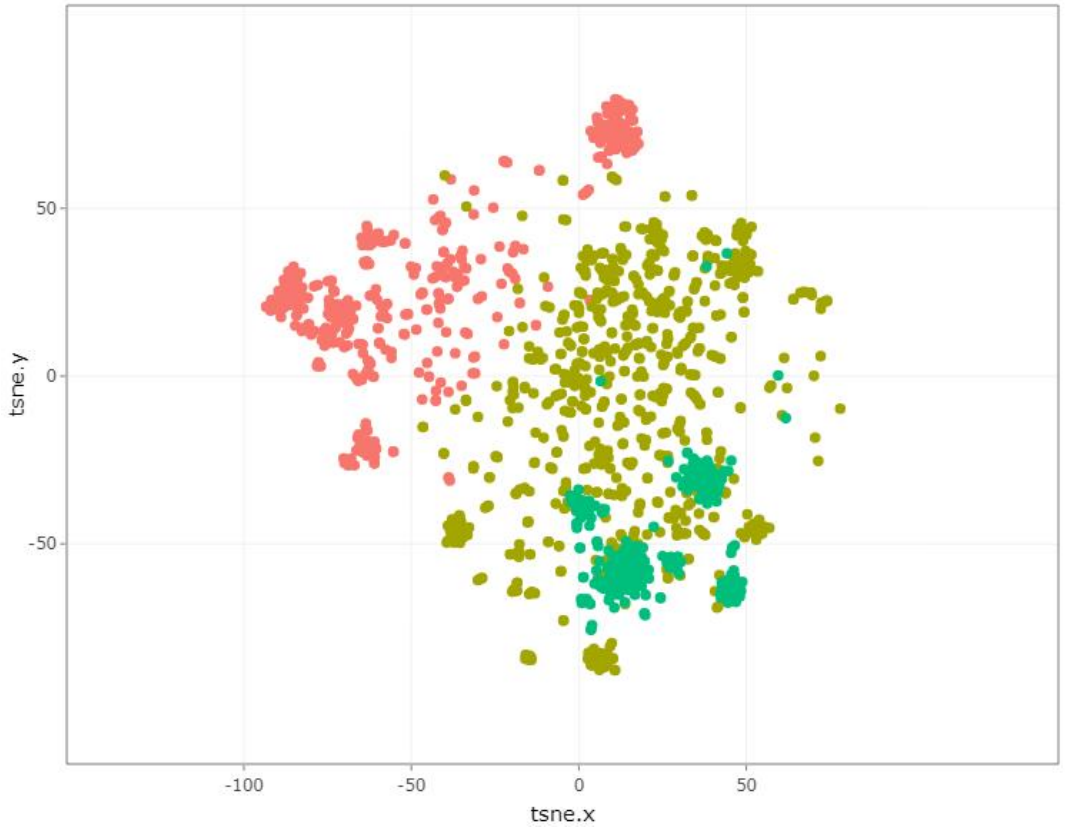
Palabras típicas: *reales, villa, ciudad, señorías*



Unigramas

Grupo	Etiqueta	N. Documentos
1	Latín y Oriental	48
2	Burocrático	508
3	Cartas deferenciales	187

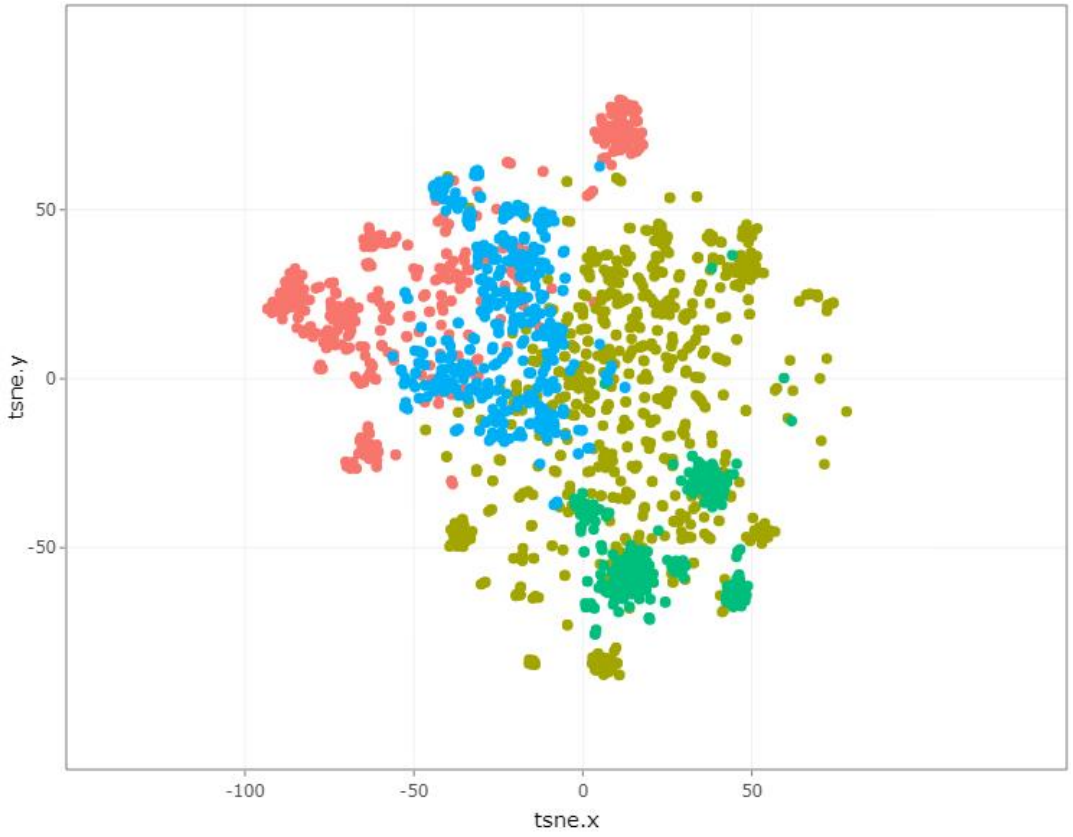
Palabras típicas: *vuestra, señoría, merced, excelencia, magestad, ilustrísima, excelentísimo*



Unigramas

Grupo	Etiqueta	N. Documentos
1	Latín y Oriental	48
2	Burocrático	508
3	Cartas deferenciales	187
4	Leyes y contratos	214

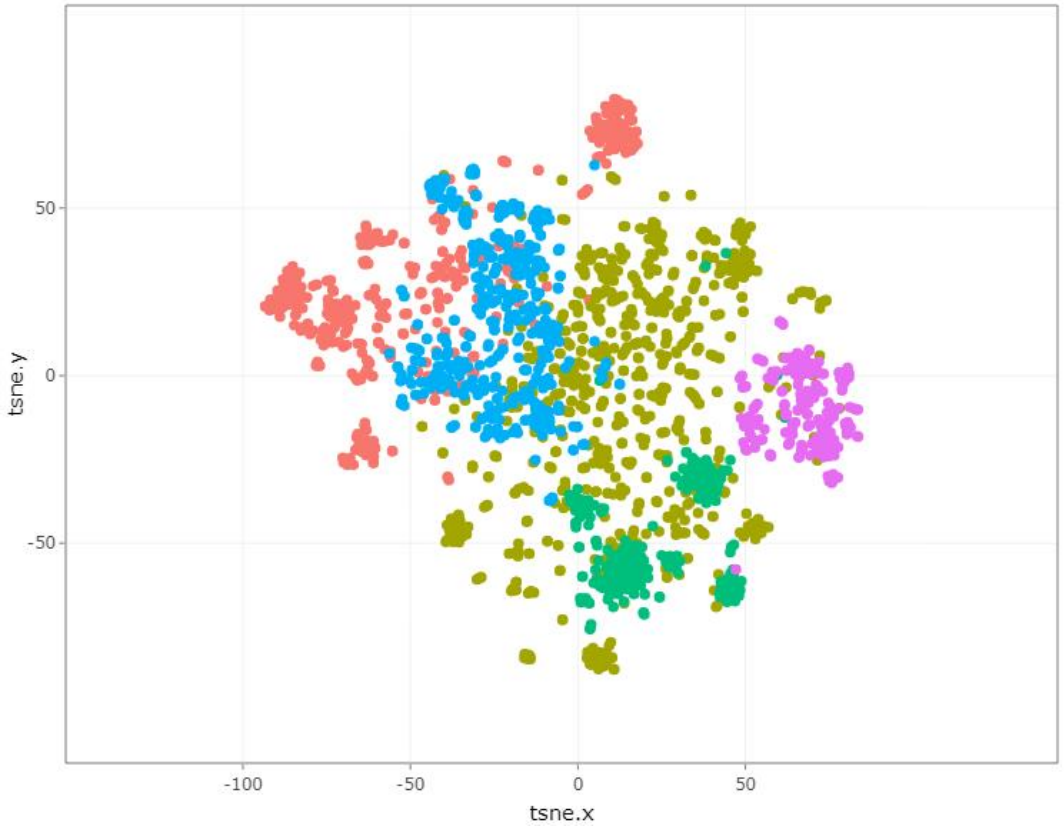
Palabras típicas: *vós, nós, monasterio, ciudad, iglesia, cabildo*



Unigramas

Grupo	Etiqueta	N. Documentos
1	Latín y Oriental	48
2	Burocrático	508
3	Cartas deferenciales	187
4	Leyes y contratos	214
5	Notas y breves	35

Palabras típicas: *niño/a, cristiano/a, llamase, agua, criatura*



Aplicaciones concretas

La complementación asindética (sin *que*)

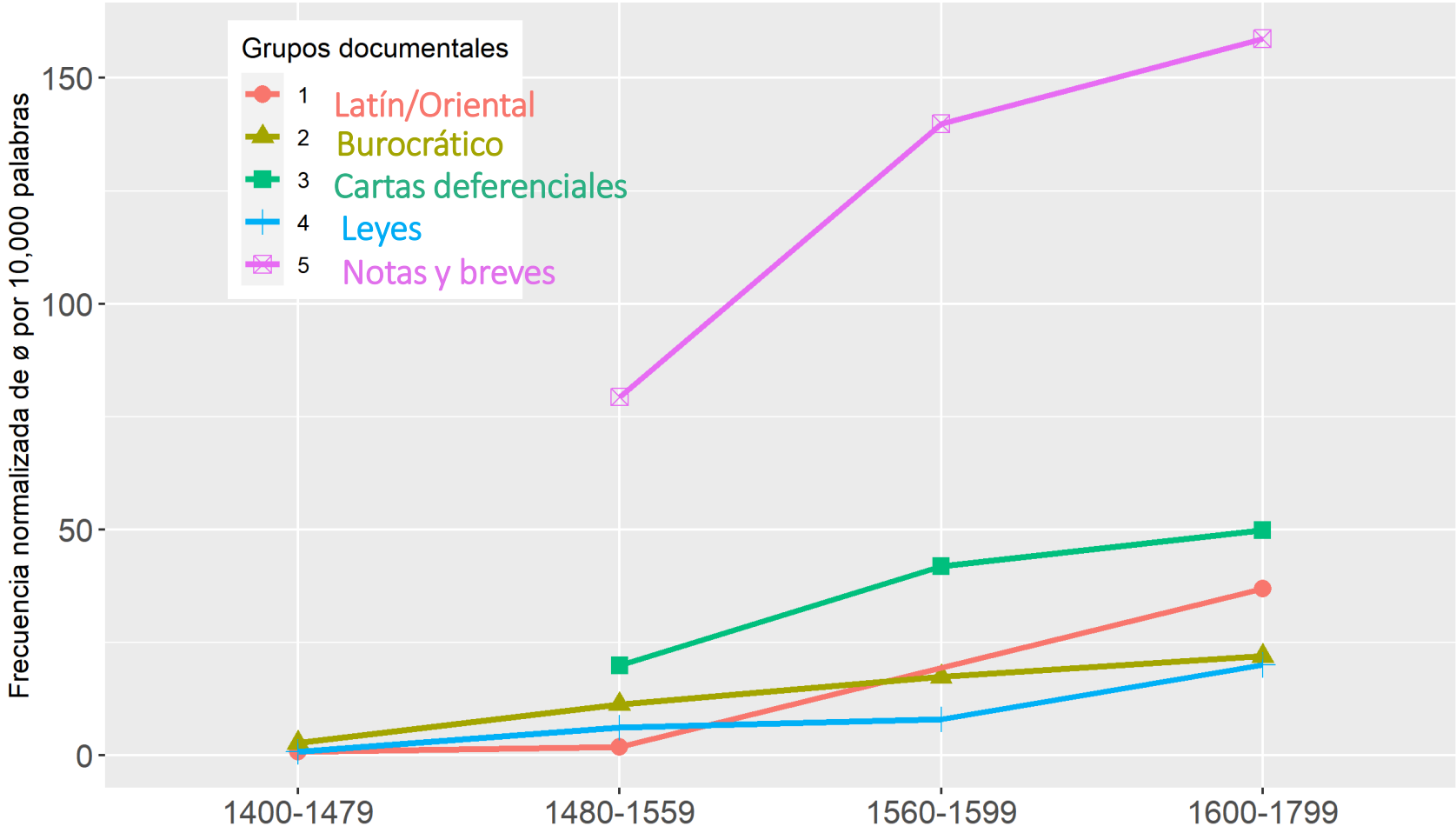
(1) ***Suplico***$_{V1}$ *a vuestra merced* **Ø *mande***$_{V2}$ *recibir essa criatura.*

(CODEA-1705, Madrid, 1596)

- Una innovación sintáctica que se desarrolla en el s. XVI (cf. Pountain 2015, Mazzola et al. *en prensa*)
- Se afianzó como un rasgo elegante de la lengua de la distancia comunicativa (Octavio de Toledo y Huerta 2011, Mazzola et al. 2020)

Grupos documentales basados en Unigramas

- Ausencia de los grupos 3 y 5 en el primer periodo
- La frecuencia de completivas asindéticas se desarrolla en estos documentos dirigidos a destinatarios prestigiosos



Skipgrams 4

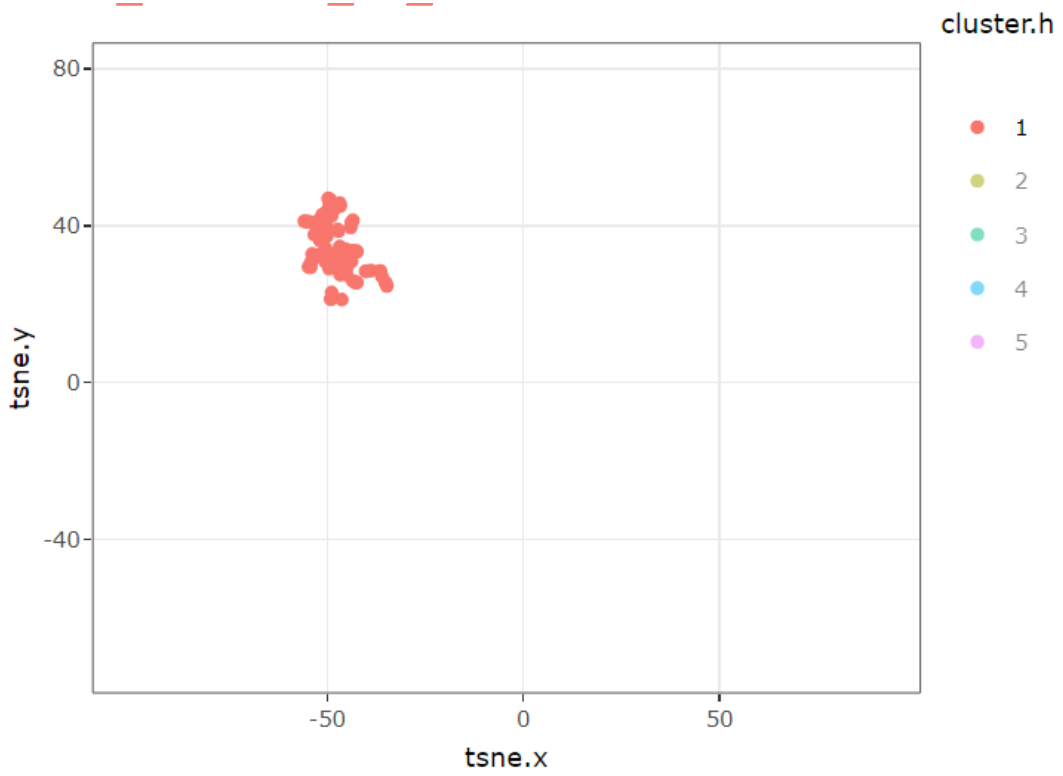
Colocaciones con elementos variables (4 palabras)

- Frecuencia mínima 10 ocurrencias

Skipgrams 4

Colocaciones típicas:
de_la_dita_iglesia,
de_la_dita_ciudat,
vós_e_los_vuestro,
e_cadaúno_de_nós

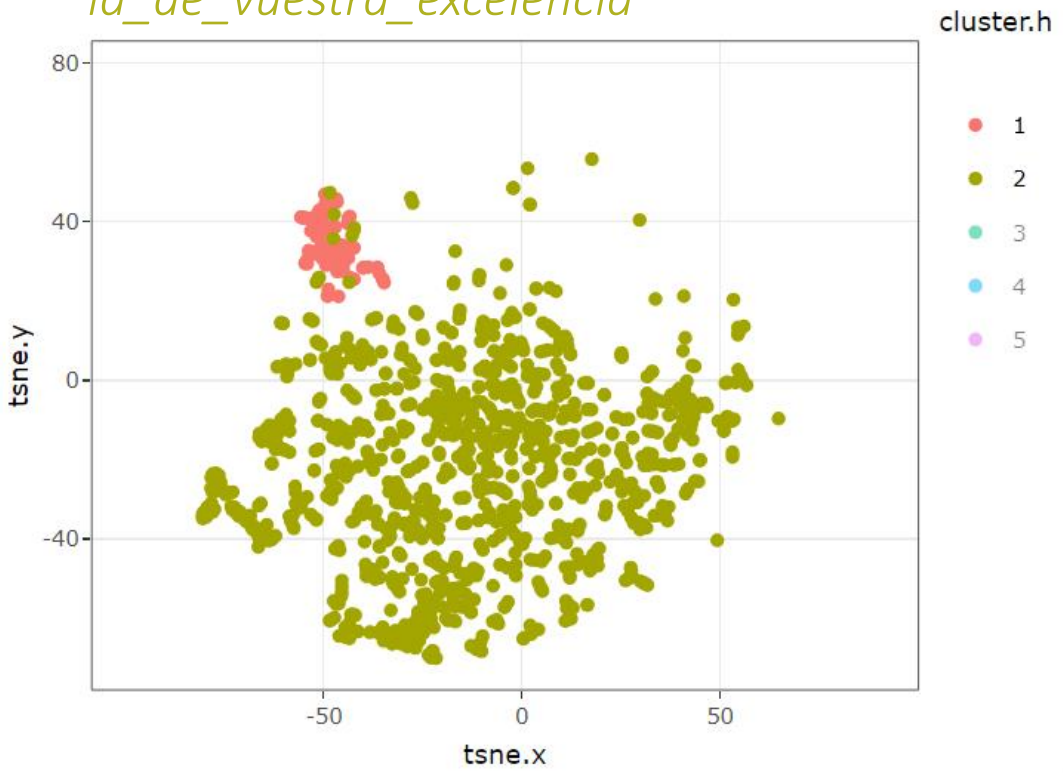
Grupo	Etiqueta	N. Documentos
1	Compraventa y testamentos orientales	49



Skipgrams 4

Colocaciones típicas:
de_la_dicha_cibdad
días_del_mes_de,
está_y_se_llama,
la_de_vuestra_excelencia

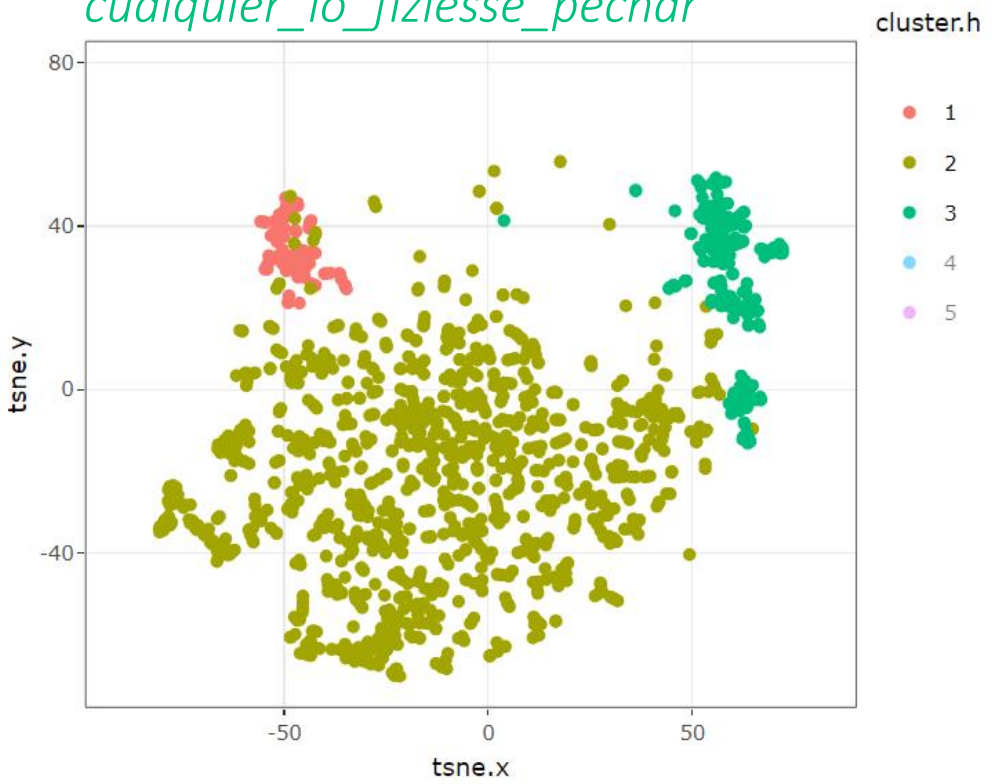
Grupo	Etiqueta	N. Documentos
1	Compraventa y testamentos orientales	49
2	Actas y Cartas	794



Skipgrams 4

Grupo	Etiqueta	N. Documentos
1	Compraventa y testamentos orientales	49
2	Actas y Cartas	794
3	Textos legislativos I	52

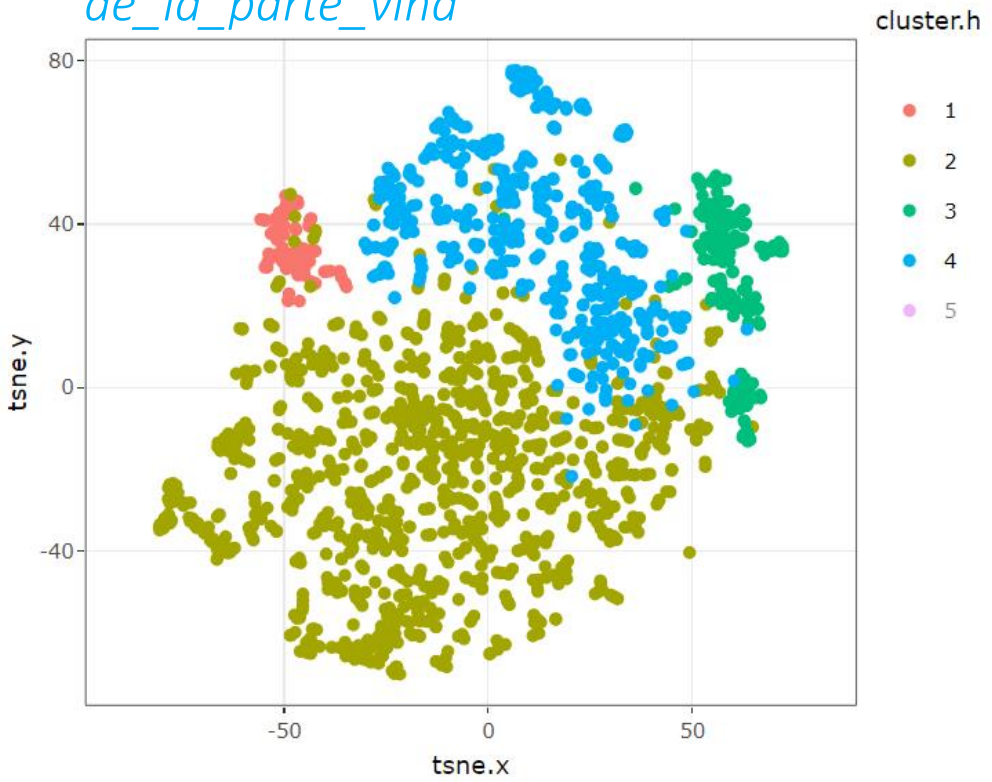
Colocaciones típicas:
del_rey_Ferrando_mio,
del_rey_confirma_Don,
nin_por_nin_por,
e_a_los_sus
cualquier_lo_fiziesse_pechar



Skipgrams 4

Colocaciones típicas:
todo_poco_e_cuanto,
In_Dei_nomine_Ego,
con_todas_suas_pertenencias,
de_la_parte_viña

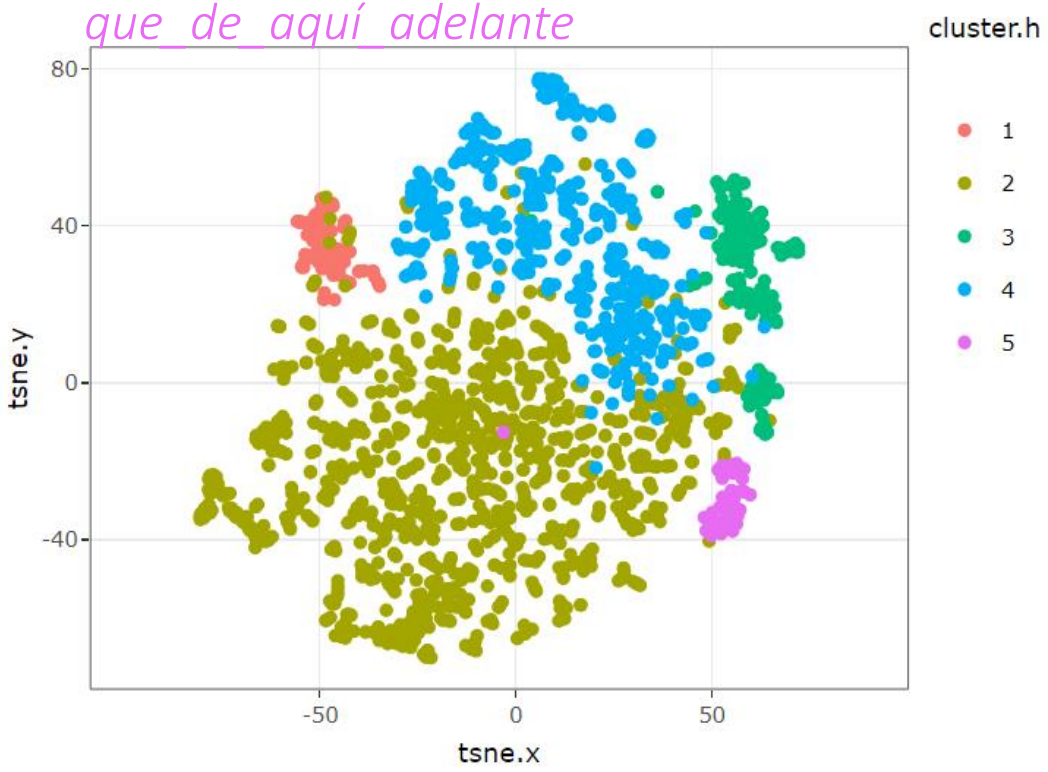
Grupo	Etiqueta	N. Documentos
1	Compraventa y testamentos orientales	49
2	Actas y Cartas	794
3	Textos legislativos I	52
4	Compraventa	104



Skipgrams 4

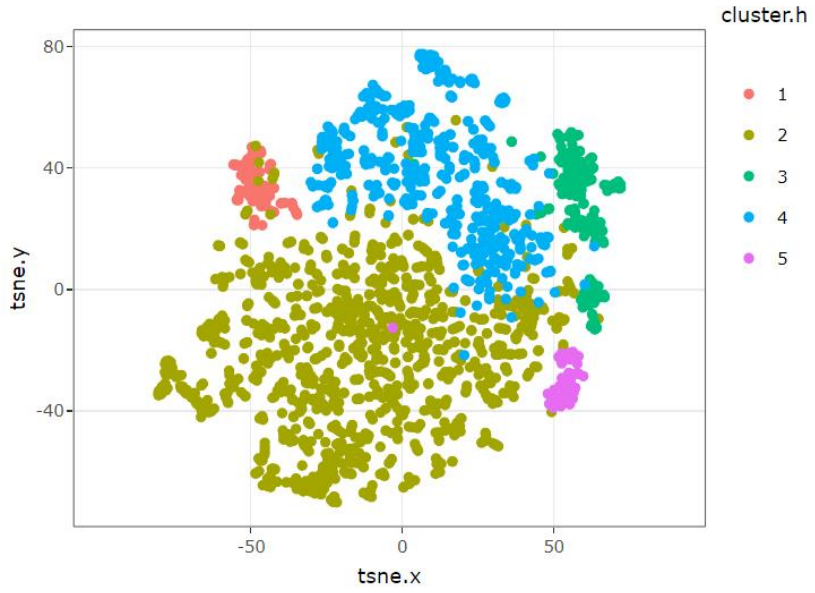
Colocaciones típicas:
ante_del_nuestro_consejo,
y_de_Tirol_etcétera,
alcalde_mayor_de_la,
de_nuestros_reinos_e,
que_de_aquí_adelante

Grupo	Etiqueta	N. Documentos
1	Compraventa y testamentos orientales	49
2	Actas y Cartas	794
3	Textos legislativos I	52
4	Compraventa	104
5	Textos legislativos II	59



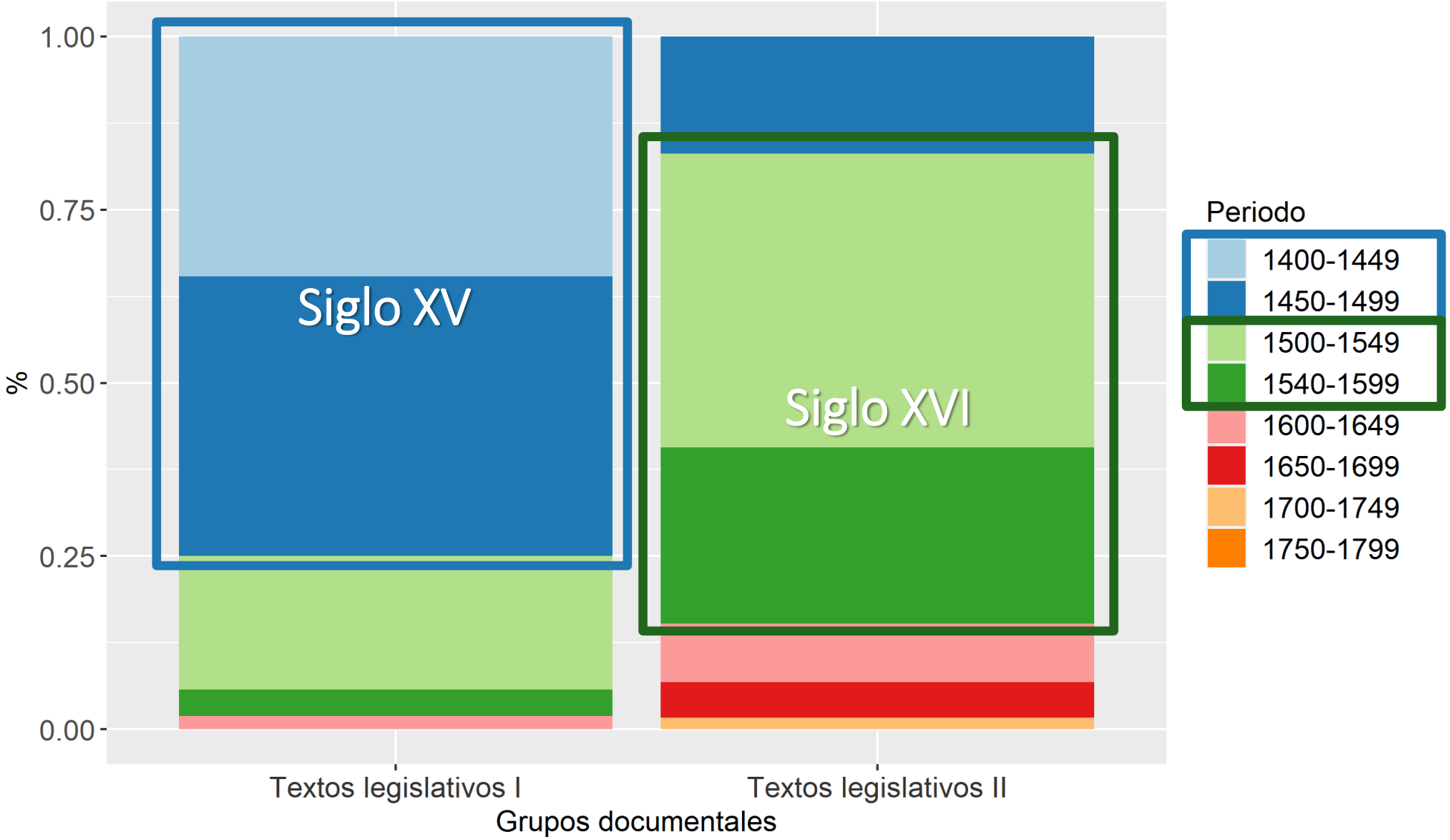
Skipgram 4

- **Diferenciación diatópica**
- grupo 1 → textos orientales
- **Diferenciación diacrónica**
- grupo 1 y 4 → solo en los siglos XV y XVI
- grupo 3 y 5 → textos legislativos de diferentes épocas y tradiciones



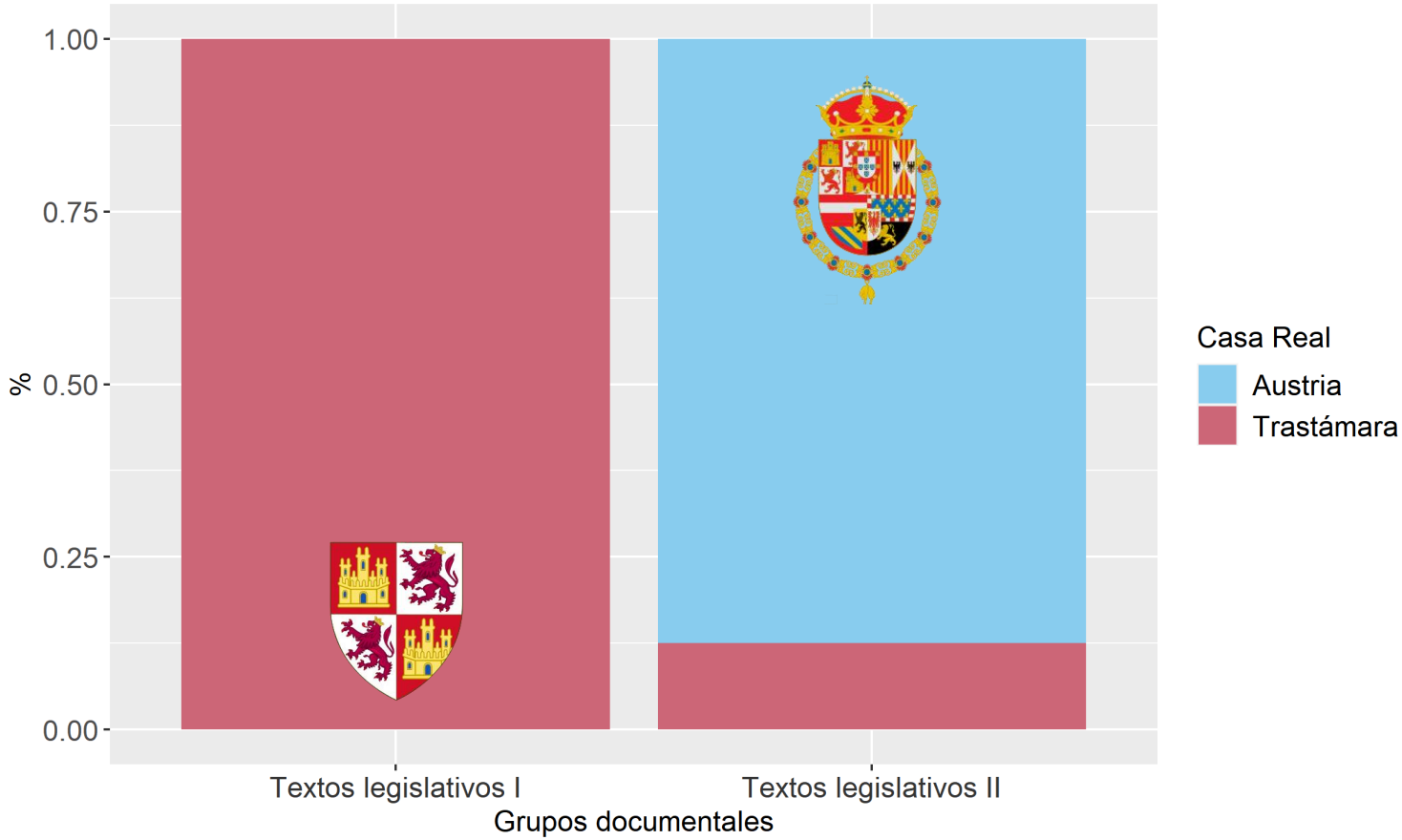
Grupo	Etiqueta	N. Documentos
1	Compraventa y testamentos orientales	49
2	Actas y Cartas	794
3	Textos legislativos I	52
4	Compraventa	104
5	Textos legislativos II	59

Skipgram 4 – textos legislativos



Skipgram 4 – textos legislativos

- A caballo entre el s. XV y el s. XVI → transición de la dinastía de Trastámara a la dinastía de Austria (Hasburgo)



Skipgram 4 – textos legislativos

- En qué consiste la diferencia?

-Los tipos de formulas

Textos legislativos I

- Formas arcaicas (*cualquier_lo_**fiziesse**_pechar*);

- referencias a miembros de la misma dinastía: *Conocida cosa sea a todos <...> cómo yo don Alfonso […] vi carta del rey don **Ferrando, mio** padre, fecha en esta guisa […]*

Textos legislativos II

- Ortografía más moderna

- los monarcas de la dinastía Austrias enumeran todos sus reinos en el íncipit de cada ley:

por la gracia de Dios rey de Castilla, de León, de Aragón, de las dos Secilias, de Jerusalem, de Navarra, de Granada, de Toledo, de Valencia, de Galizia, de Mallorcas, de Sevilla, de Cerdeña, de Córdova, de Córcega, de Murcia, de Jaén, de los Algarves, de Algezira, de Gibraltar, de las islas de Canaria, de las Indias, islas e tierra firme del mar Océano, condes de Varcelona, señores de Vizcaya y de Molina, condes de Flandes y de Tirol, etcétera

Skipgram 4 – textos legislativos

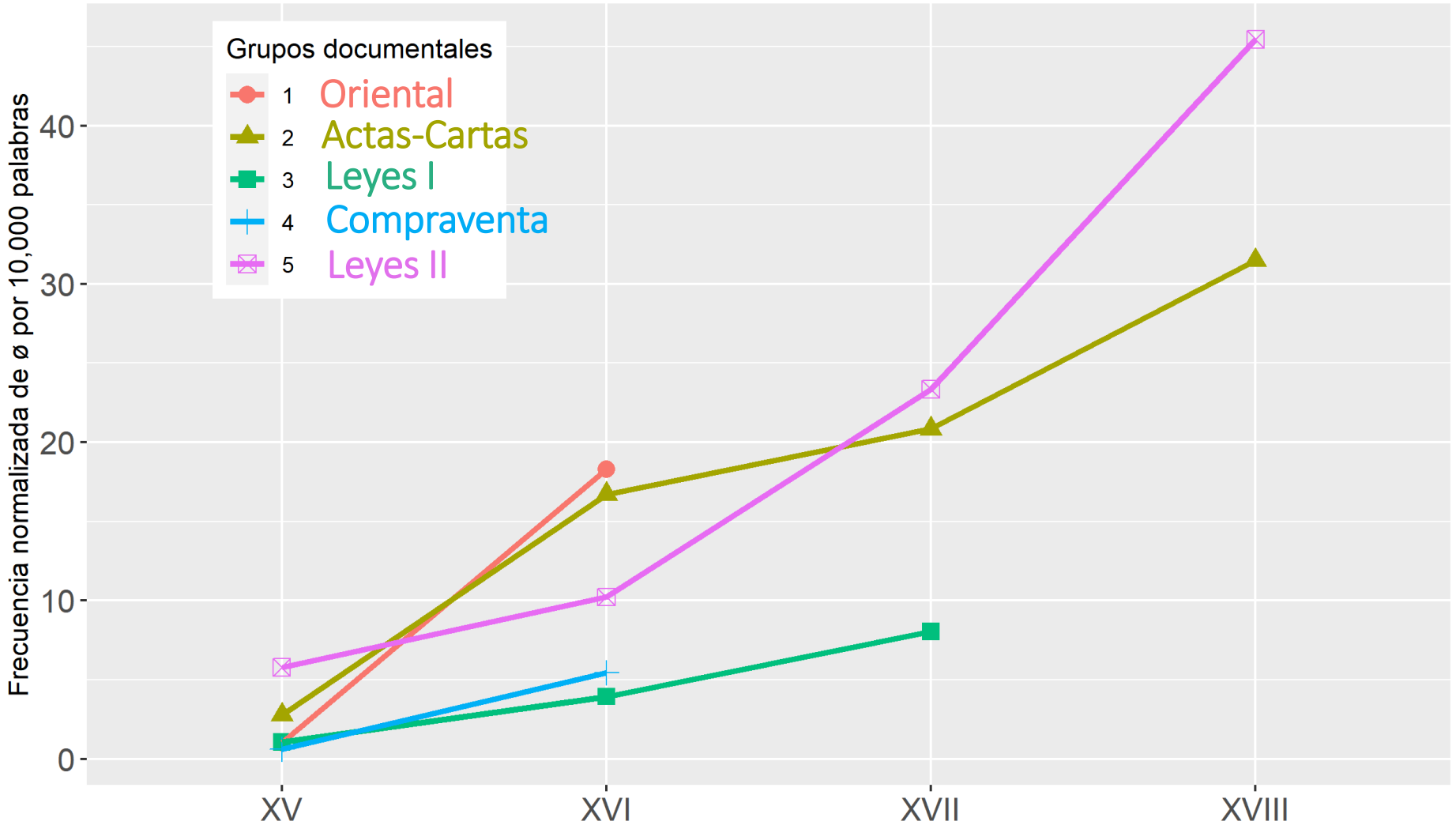
Se necesita un análisis textual más minucioso para averiguar estas observaciones y profundizar el tema

"todo esto no es cuestión de frecuencia o de datos cuantitativos" (Kabatek 2013: 20)

→ Las frecuencias y los datos cuantitativos pueden ser una poderosa herramienta en una primera fase de análisis

Grupos documentales basados en skipgrams 4

- Marginal en el s. XV
- En el s. XVI en la clase de Actas y Cartas
- Los ejemplos más tardío frecuentes en la nueva clase de textos legislativos
- Correlación entre una variante sintáctica innovadora y TDs nuevas?



Conclusiones

- La técnica aquí presentada tiene la ventaja de utilizar los elementos que constituyen un texto para medir semejanzas y agrupar textos
- puede ser un acercamiento interesante para llegar a la identificación de TD, por lo menos a un nivel de generalización bastante grueso.

Ventajas

Clasificación textual basada en los elementos que componen el texto

- Detecta TDs diatopicamente y diacronicamente marcadas
- Nos mostró que:
  - Hay TDs más homogéneas que otras, que se corresponden con tipología textuales clásicas
  - Hay TDs menos homogéneas, que pueden abarcar distintos tipos textuales
- Las aplicaciones de esta clasificación dan resultados interesantes

- los cambios lingüísticos se actúan a través la mediación de ciertas tradiciones discursivas (Cartas y Notas, y Textos legislativos más modernos)

Desventajas

- Interpretación de los grupos: no puede ser independiente de las clasificaciones tradicionales, que se utilizaron también para definir y entender cada grupo
- Muchísimas posibilidades combinatorias de los parámetros, que pueden llevar a clasificaciones diferentes (numero de grupos, criterios…)
- Textos enteros → ¿cómo integrar la idea de zonas discursivas?
- Elegir los criterios de agrupación:

    - Unigramas vs. Skipgram: los resultados son diferentes

- ¿Cuál criterio es el más adecuado?
- ¿Hasta qué punto podemos decir que esta metodología detecta TDs?

Referencias

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Mark Davies, James K. Jones & Nicole Tracy-Ventura. 2006. Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora* 1(1). 1–37. https://doi.org/10.3366/cor.2006.1.1.1.

De Pascale, S. 2019. *Token-based vector space models as semantic control in lexical lectometry*. Doctoral Dissertation. https://lirias.kuleuven.be/retrieve/549451 (22 January, 2021).

GITHE. 2015. Codea+ 2015. *Corpus de documentos españoles anteriores a 1800*. http://www.corpuscodea.es/ (10 January, 2019).

Kabatek, Johannes. 2005. Tradiciones discursivas y cambio lingüístico. *Lexis: Revista de lingüística y literatura* 29(2). 151–177.

Kabatek, Johannes. 2013. ¿Es posible una lingüística histórica basada en un corpus representativo? *Iberoromania* 77(1). https://doi.org/10.1515/ibero-2013-0045. https://www.degruyter.com/doi/10.1515/ibero-2013-0045 (1 July, 2020).

Kabatek, Johannes, Philipp Obrist & Valentina Vincis. 2010. Clause linkage techniques as a symptom of discourse traditions: Methodological issues and evidence from Romance languages. In Heidrun Dorgeloh & Anja Wanner (eds.), *Syntactic Variation and Genre* (Topics in English Linguistics), vol. 70, 247–276. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110226485.2.247. https://www.degruyter.com/view/books/9783110226485/9783110226485.2.247/9783110226485.2.247.xml (21 April, 2020).

Mazzola, Giulia, Bert Cornillie & Malte Rosemeyer. En prensa. Asyndetic complementation and referential integration in Spanish. A diachronic probabilistic grammar account. *Journal of Historical Linguistics*.

Mazzola, Giulia, Bert Cornillie, Malte Rosemeyer & Stefano De Pascale. 2020. Socio-stylistic aspects of syntactic variation: the case of Spanish asyndetic complementation between the 15th and the 18th century. Conference presented at the 53rd Annual Meeting of the Societas Linguistica Europaea. https://osf.io/fe6tc/.

Octavio de Toledo y Huerta, Alvaro. 2011. Santa Teresa y la mano visible. Sobre las variantes sintácticas del Camino de perfección. In Mónica Castillo Lluch & Lola Pons Rodríguez (eds.), *Así se van las lenguas variando: nuevas tendencias en la investigacíon del cambio lingüistico en español* (Fondo Hispánico de Lingüística y Filología v. 5), 241–304. Bern ; New York: Peter Lang.

Pountain, Christopher. 2015. Que-deletion: the rise and fall of a syntactic fashion. In Francisco Dubert García, Gabriel Rei-Doval & Xulio Sousa (eds.), *En memoria de tanto miragre. Estudos dedicados ó profesor David Mackenzie*, 143–159. Universidad de Santiago de Compostela: Servizo de Publicacións e Intercambio Científico.

¡Gracias!

KU LEUVEN

giulia.mazzola@kuleuven.be
stefano.depascale@kuleuven.be