Appliying Topic Modeling Techniques to Degraded Texts

Spanish Historical Press during the *Transición* (1977-1982)

Carlos G. Figuerola

University of Salamanca, 2018

Text Mining and Historical Press

- Topic modeling and another text mining techniques are used often with historical texts
- This makes sense if we work with a huge amount of text
- Digitalized Newspapers is a useful historical source

Some problems in digitalizing newspapers

- Often, digitalizing process consists on passing trough a scanner the pages of newspapers
- This requires an OCR software to be able to access to text
- OCR has to deal with old typefaces, degraded paper and templates of pages
- The aim of this work is to check if those techniques perform well with fuzzy and noisy texts

The ABC newspaper

- we choose the ABC newspaper to experiment with *toping modeling* techniques
- ABC began in 1903
- All numbers since then are fully digitized.
  - digitalization consists in passing pages through a scanner, and then trough an OCR software.
- There we have templates in columns, several typographical fonts, etc.

Los otros dos ex carcelados son Diego Elorrieta Derio, de cincuenta y cinco años de edad, casado, con una hija. Fue condenado en septiembre de 1973 a doce años de prisión, y Pedro Aulestiaondarroa, natural de Ondárroa, de veinte años de edad, estudiante, que fue condenado el 18 de julio de 1975. Este detenido había cumplido a lo largo de su estancia en prisión ciento cuarenta días en celda de castigo.

Continúan saliendo reclusos

Con la aplicación de las nuevas medidas, en la noche de ayer eran ya 296 los presentencias.

Siempre según estas mismas fuentes, próximas a los abogados defensores, en el caso de que la amnistía no sea aplicada de oficio por los Tribunales Militares, los interesados pedirán a la Audiencia Nacional que las aplique.

Con relación a los que se encuentran en prisión por estar relacionados con los secuestros, no está claro si se podrá aplicar la amnistía, al haber comenzado éstos antes de la fecha del referéndum; pero, de no ser estimado así, los abogados recurrirán al Tribunal Supremo.

EL DOMINGO

"IZQUIERDA DEMOCRATICA" Y LA "F. P. D." PUEDEN FIRMAR UN PACTO DE FEDERACION

Madrid. (De nuestra Redacción.) La «Federación Popular Democrática» e «Izquierda Democrática» pueden llegar a la constitución de una federación de partidos, con presidencia, secretariado general y ejecutiva comunes. Esta federación tendría carácter transitorio, en tanto los respectivos congresos, y posteriormente un Congreso Constituyente, decidan sobre la unidad de ambos partidos demócrata cristianos.

La confirmación de la unidad se retrasará hasta después de las elecciones, por considerar «Izquierda Democrática» que no queda tiempo material para realizar todo el proceso de integración con anterioridad a los comicios. Sin embargo, desde el momento mismo en que se apruebe el acuerdo de federación entrarán en funcionamiento los órganos comunes y de hecho funcionará como un único partido.

Por el momento, según fuentes democristianas, ha sido elaborado ya un documento base en el que se regula el funcionamiento de los órganos comunes transitorios.

EL «P. P. D. C.».—La operación de unidad podría extenderse igualmente hacia el «Partido Popular Demócrata Cristiano» y la «Unión Democrática Española». De hecho, el primero de estos partidos acudió a la reunión conjunta de las ejecutivas de «Izquierda Democrática» y «Federación Popular Democrática», pero su representante en la reunión, Iñigo Cavero, no tenía facultad decisoria y asistió como observador.

El «Partido Popular Demócrata Cristiano» mantiene paralelamente conversaciones orientadas a la fusión con «Unión Democrática Española». Por el momento se han redactado varios borradores de trabajo, pero no parece haberse concretado nada. Fernando Alvarez de Miranda, consultado por A B C, se remitió al comité político que se reúne el próximo domingo en Oviedo, y en el que se informará de las gestiones realizadas hasta el momento.

FIRMA DEL PACTO.—El próximo domingo, en una reunión que mantendrán en el Colegio Mayor San Pablo el Comité federal del «F. P. D.» y la ejecutiva de «Izquierda Democrática», se espera que se firme el acuerdo de federación y el compromiso de convocar sus respectivos Congresos para aprobar la integración.

Anteriormente, el sábado, se reunirá el Comité federal de la «Federación Popular Democrática» para debatir y aprobar el pacto de federación.

ABC after OCR

SANCHO D AVILA, 19 - T«!s, 245 80 43 - 246 20 9 8
El «Partido Popular Demócrata CristiaMADRID-28
no» mantiene paralelamente conversaciones orientadas a la fusión con
«unión Democrática Española». Por el momento se
han redactado varios borradores de trabajo, pero no parece haberse
concretado líada. Fernando Alvarez de Miranda, consultado por ABC, se
remitió al comité político que se reúne el próximo domingo en
Oviedo, y en el que se Informará de las
gestiones idealizadas hasta el momento.
FIRMA DEL PACTÓ.—El próximo doa diez minutos de la nieve, edificable
mingo, en una reunión que mantendrán en
Precio total: 747.000 pesetas
el Colegio Mayor San Pablo el Comité federal del «P.P.p.» y la ejecutiva
de «IzFACILIDADES
quierda Democrática», se espera que se firme el acuerdo de federación y
el comproTeléfonos 253 56 64 y 254 58 86
miso de convocar sus respectivos Congresos para aprebar la Integración.
Anteriormente, el sábado, se reunirá el P I S O E N F U B N T E R R A B
1 A
GóMt! ÍNléral de la «Badéraclóa Popular ©áÜício Mlramar Torreón, todo
exterior.
232 m2 útiles. Info»na<íloa en Madrid,
D^b&ráliél^ Q$m debatir y aprobar tí
Teiéfoiio 21$ 28 77. Tardes.
pací» dé fgáeráclón.

El País as a benchmark

- El Pais newspaper began in 1977
- it has a digital archive,
  - made by transcribing by hand the individual news
  - no scanner neither OCR
  - navigable trough the internet

The Transición period

- it goes from 1977 to 1982
- we have full archives from ABC and El País for this time period
- comparison between both newspaper can tell us about the reliability of applications of those techniques on degraded texts (ABC)

Topic Modeling

- Latent Dirichlet Allocation is the most usual
  - in a collection of documents are a set of topics
  - each document has an specific amount(%) of each topic
  - LDA produces a set of terms defining each of topics
  - also, the percentage of every topic in every document
- we have easy tools to apply LDA (Mallet and others)

Topics from El País

agua aceite productos agricultura agricultores medio enfermedad zonas	water oil products agriculture farmers environment sickness areas
bank chemical valores índice mercado steel ltd bolsa	bank chemical values index market steel ltd stock-exchange
política economía precios crisis crecimiento empresas mercado inflación	policy econonomy prices crisis increase enterprises market inflation
soviética unidos urss israel moscú guerra china paz	soviet united ussr moscow war china peace
millones pesetas crédito empresa economia bancos dinero valor	milions pesetas credit enterprise economy Banks money value
internacional presidente unidos argentina militar carter reagan británico	international president united argentina military carter reagan british
trabajadores huelga empresa economia convenio sindicatos paro acuerdo	workers strike enterprise economy unión-agreement trade-unions unemployment agreement
francia gobierno europa relaciones internacional exteriores otan francés	france government europe relationship international foreign nato french
petróleo energía nuclear dólares producción mercado industria precio	petrol energy nuclear dolars production market industry price

Exotic topics from ABC

interna necesita niños casa sueldo informes cocina teléfono llamar madrid playa	domestic-maid needed children home salary reports kitchen phone call
mercedes estrenar extras bmw particular completo nuevo automóviles	mercedes brand-new extras bmw private full new cars
metros dormitorios piso garaje piscina facilidades zona particular	meters bedrooms flat garaje swimming-pool easy area private

Classified ads in ABC

GENERALISIMO, vacío, amueblado estrenar. 2503348.
ARGÜELLES, amueblado vacío, confortabilísimo. 2504295.
OFICINAS con teléfono. 2500942.
SALAMANCA, amueblado, desamueblado, elegante. 2595840.
¡¡¡**AGUAYO**!!! English Française, apartamentos chalet, oficinas, pisos. 2500942.
DIRECTAMENTE. Apartamentos Rueda, lujosísimos. Zona Castellana. Servicios incluidos. Teléf. 4469777.
ALMAGRO, 10. Apartamentos «Andrómeda». Estrenar, climatizados, alquileres semanales-mensuales, todos servicios. 4470900.
PRECIOSOS apartamentos climatizados, zona glorieta Bilbao. Servicio. Alquileres semanales - mensuales. Velarde, 9. 2328920.
CASTELLANA, 56. Importante piso, 90.000, dos plazas garaje, 400 metros. 2223133.
ROSALES, 28, oficinas, 50.000, superconfortabilísimas. 2224833.
RESIDENCIAL, chalet, 85.000. Aravaca. 2319290.
AMUEBLADO, 40.000, superconfortable. 2320644.
RESIDENCIAL, vacío, 40.000. 2223133.
HABANA, 140.000. Oficinas, quince habitaciones, 470 metros. 2500942.
ARAPILES, 40.000, amueblado, cuatro dormitorios. 2503348.
CIUDAD Periodistas, 35.000, vacío, 220 metros.
junto a plaza de España. Facilidad de aparcamiento. Directamente propietario. Teléfonos 2478572 y 2486826.
MONTERA. Pisos, apartamentos oficinas residenciales, vacíos amueblados. 2315690.
ALQUILAMOS oficinas zona Cibeles-Colón. Teléfono, aire acondicionado. Información. 2759190.
SE alquila oficina 330 metros, planta primera, edificio moderno, con teléfono. Calle General Martínez Campos, 9. Información: Teléfono 4552102.
APARTAMENTOS alquiler, amueblados, estrenar, garaje, piscina, aire acondicionado. Alberto Alcocer, 43.
APARTAMENTOS amueblados zona Castellana, servicios parking, sauna, gimnasio, cafetería. Edificio Cónsul. Alquilo directamente. Señor Carro. 4191650.
OFICINA amueblada nueva, 4 despachos. María de Molina, 26. Mañanas.
DISPONEMOS oficinas todas zonas, superficies, precios. 2312324.
DIVERSIDAD pisos, chalets, oficinas disponemos. Inmediata ocupación. 2326553.
SOL. Mayor, 1, segundo. Oficinas y locales.
APARTAMENTOS semanales desde 4.200, mensuales desde 16.800, servicios incluidos. Vargas, 3 (esquina General Sanjurjo) 4412500.
OFICINAS. Salesas. Ocho
PUERTA Hierro, alquilo piso amueblado, 4 dormitorios, 3 baños, salón, comedor separado, despacho, hermosa vista, octava planta, Arroyo Fresno, 26, cerca hotel Monterreal, razón portería. 2793552.
APARTAMENTOS amueblados de lujo, zona plaza España, espaciosos, luminosos, climatizados, TV., servicios completos incluidos. Véalos: Onésimo Redondo dieciséis. 2425900.
OFICINAS, teléfono, garaje propio, laborables Cea Bermúdez, 14.
OFICINAS, inmejorables zonas. 2623726. Baikal.
GENERALISIMO, vacío, lujo, cuatro dormitorios, garaje, orientación Mediodía. 60.000. 6501498.
PEQUEÑO, amueblado lujosísimo, gran terraza, chimenea. 2599521.
¡¡¡**OFICINAS**!!! Desde 150 m2. hasta edificios de 15.000 m2. Exclusivamente zonas residenciales, Uve y Eme «Boutique Inmobiliaria». Castelló, 41. Teléfonos 2755511 - 62.
APARTAMENTO junto Puerta del Sol, habitación, salón amplísimo, cocina, baño, dos terrazas más invernadero, de 10 m2 propio pintores, escultores artesanos. 2667705.
APARTAMENTO con garaje. 31.000. 4197101.
ATICO, Salamanca, cuatro dormitorios. 4488174.
PISO Salamanca. 4022838.
¡**OFICINAS**! General Mola, 100 metros, 22.000
cuadrados, 27.000 pesetas. Teléfono 248 11 81. De cuatro a seis.
OFICINA céntrica moderna, teléfono, aparcamiento, Cuesta Santo Domingo, 11.
OFICINA, 80 metros, mejor tramo paseo Habana, dos amplios despachos, calefacción, servicio. 25.000. 4469801.
APARTAMENTO amueblado, estrenar, próximo paseo Habana, calefacción central. 34.000. 4469802.
CHAMARTIN, vacío, lujo, dos dormitorios, calefacción, agua centrales, teléfono, garaje, 28.000. 4469801.
EDIFICIO oficinas, avenida América, teléfono, 90 metros, 30.000. 2423616.
FLORIDA, chalet amueblado estilo inglés, 140.000. 2422542.
PUERTA Hierro, amueblado, estrenar, muy lujoso, 70.000. 2422542.
MEDICO, consulta lujo, totalmente equipada. Barrio Salamanca. 55.000. 2152342.
OFICINA Viso, 200 m. Renta interesante. 2419957.
VACIOS amueblados Argüelles, Generalísimo, Salamanca, todas rentas. 2419957.
CHALETS Chamartín, Aravaca, Somosaguas. Todas rentas. 2419957.
OFICINAS chalet Chamartín, 400 m., buena renta. 2419957.
OFICINAS Generalísimo, Argüelles, Salamanca. Todas superficies, todas rentas. 2419957.

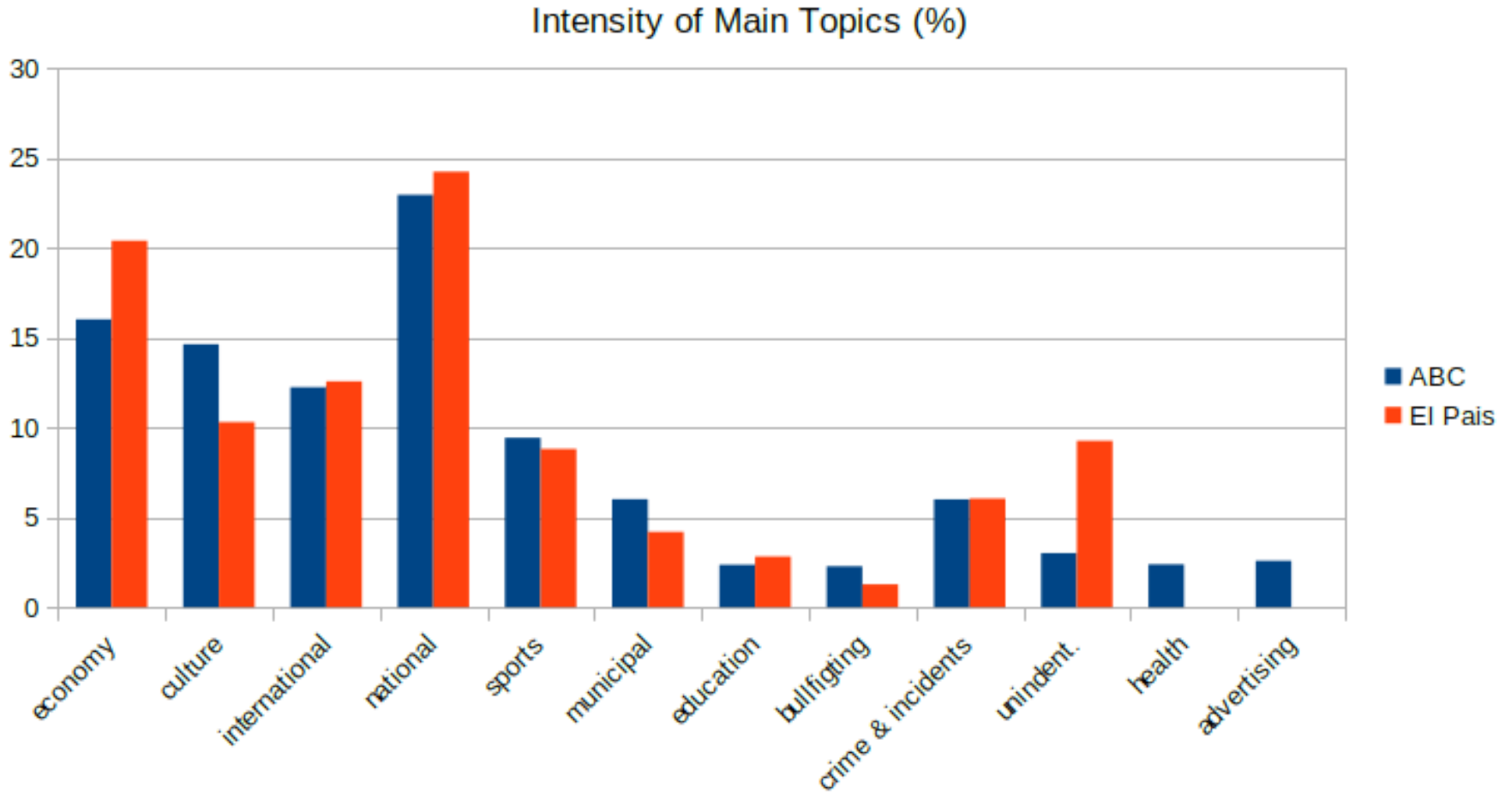
Topics from ABC (2º round)

puerto mar avión zona aguas aeropuerto costa pesca acuerdo	harbour sea plane area waters airport coast fishing agreement
millones pesetas economía mercado empresas banco dólares productos	milions pesetas economy market enterprises bank dollars products
trabajadores huelga empresa acuerdo reunión sindicales comisión gobierno	workers strike enterprise agreement meeting trade-union comission government
energía pesetas precios producción consumo toneladas mercado nuclear	energy pesetas prices production consumption tons market nuclear
banco junta consejo pesetas accionistas administración sociedad acciones	bank board council pesetas shareholders administration society stock-options
unidos soviética presidente moscú gobierno carter internacional militar	united soviet president moscow government carter international military
israel presidente ministro gobierno irán paz árabes acuerdo	israel president ministry government iran peace arabes agreement
gobierno partido ministro presidente francia parís elecciones socialista europa	government party ministry president france parís elections socialist europe

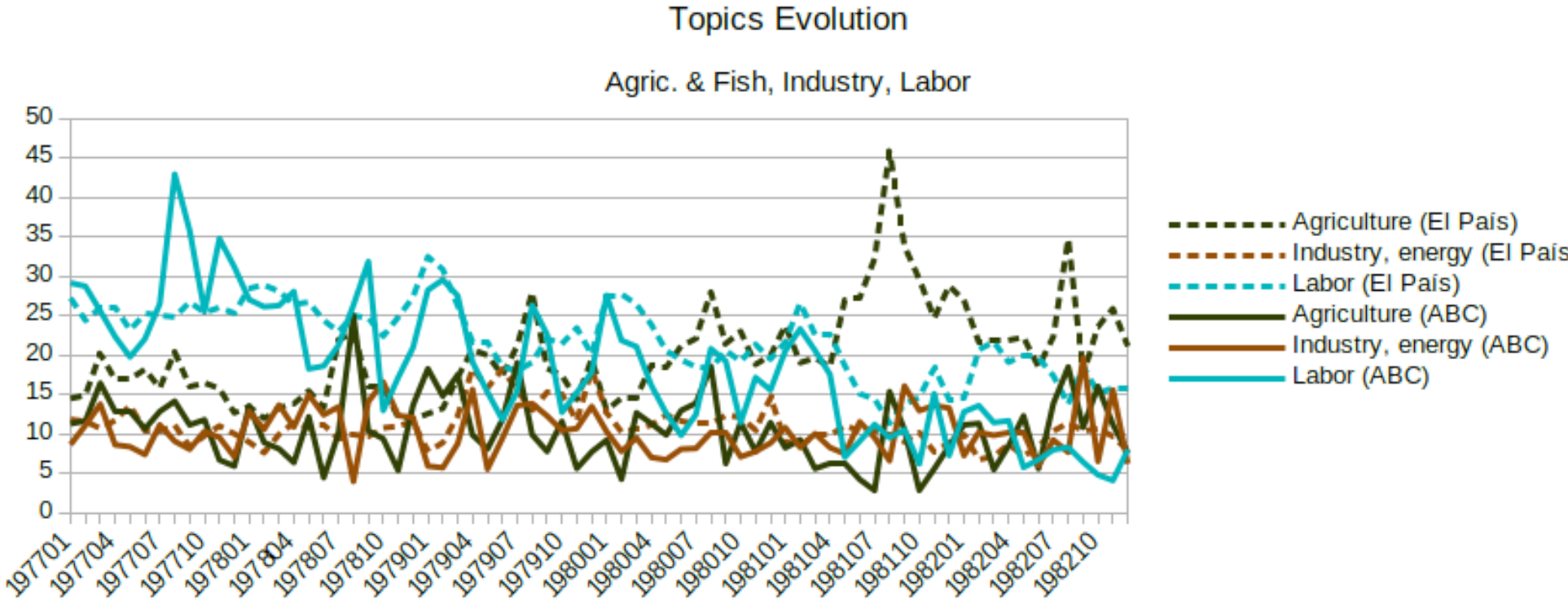
Some figures

	El País	ABC	ABC (2º round)
1977	41359	33963	15967
1978	40105	32831	14574
1979	41396	33084	14419
1980	42150	33491	14479
1981	40090	33091	14248
1982	40224	37096	16303
Total	**245324**	**203556**	**89990**

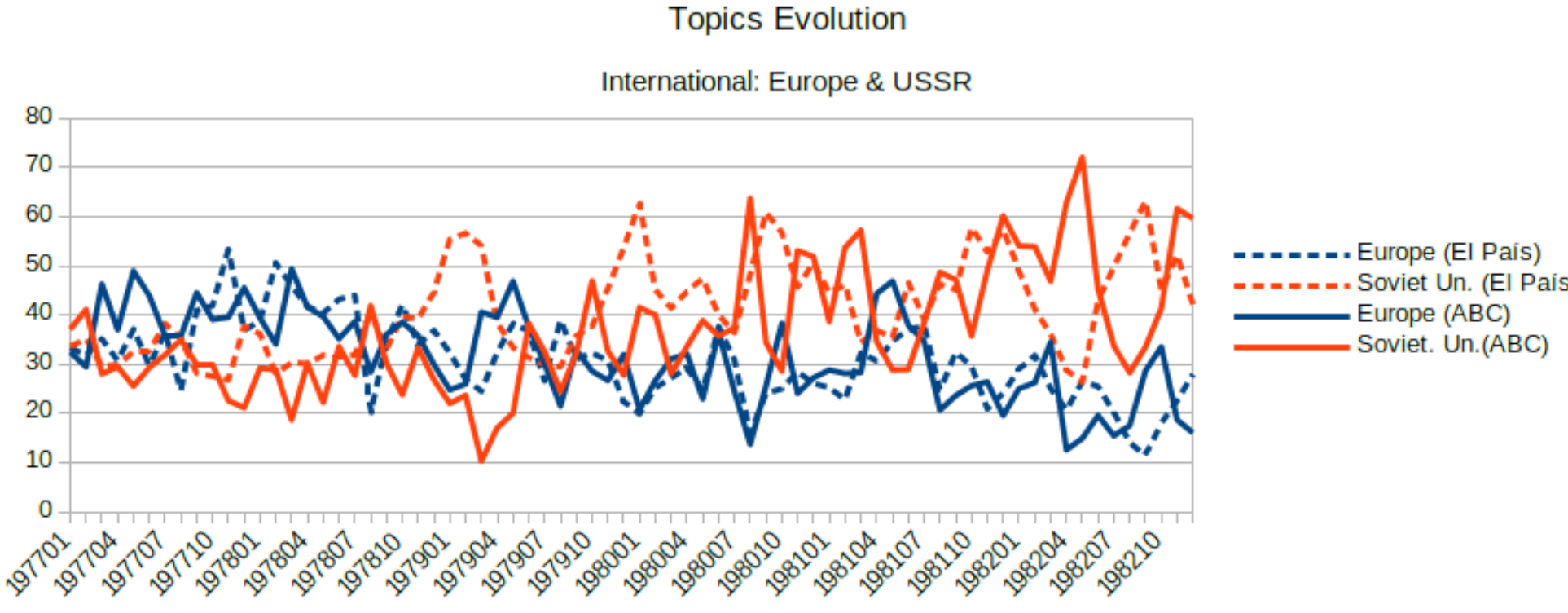
Intensity of Topics



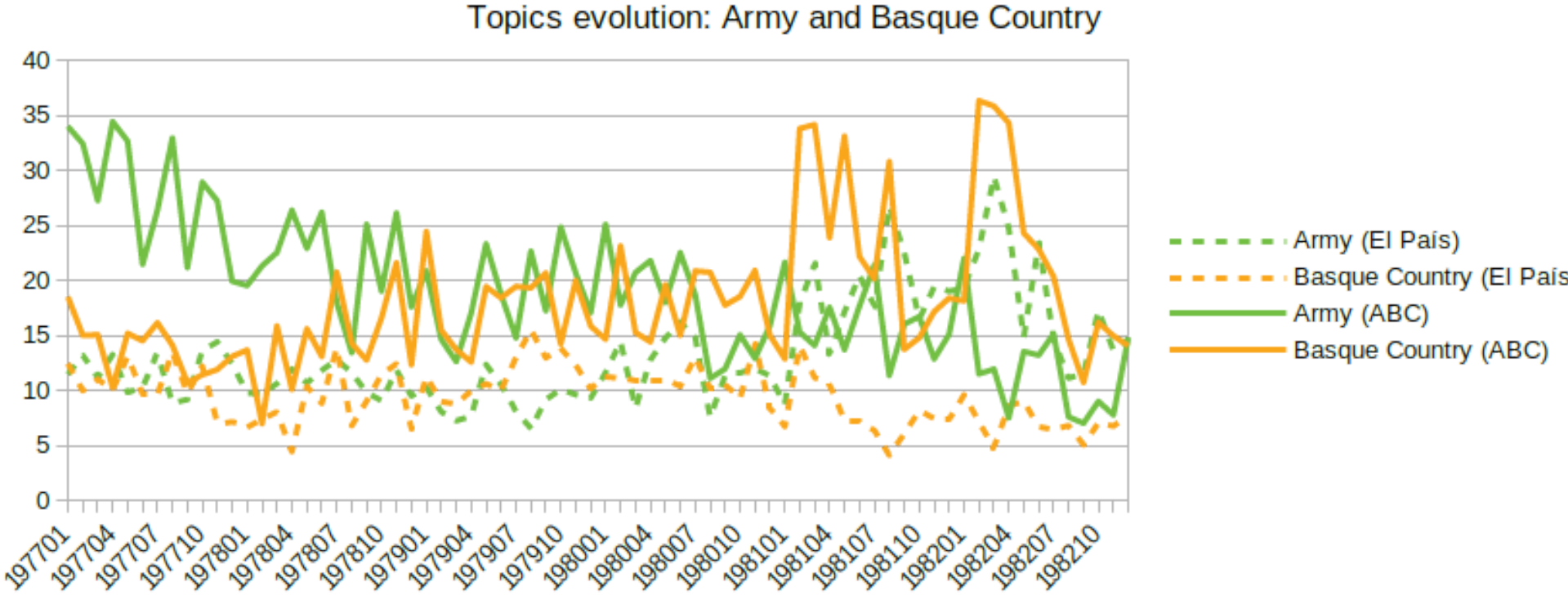
Economy


Topics Evolution
Agric. & Fish, Industry, Labor
50
45
40
35
30
25
20
15
10
5
0
197701
197704
197707
197710
197801
197804
197807
197810
197901
197904
197907
197910
198001
198004
198007
198010
198101
198104
198107
198110
198201
198204
198207
198210
Agriculture (El País)
Industry, energy (El País
Labor (El País)
Agriculture (ABC)
Industry, energy (ABC)
Labor (ABC)

International



Army and Basque Country



Conclusions

- we have applied LDA to noisy text comming from scanned newspaper
  - full page as analysis unit
  - irregular templates (columns, several widths, etc.)
  - old type-faces, advertising, …
- we applied LDA also to text comming from another newspaper
  - plain and clean text
  - individual, single news are the analysis unit
  - no advertising

Conclusions

- topics are the same for both newspapers
  - although the terms for some of topics are different
- evolution of intensity in topics shows, in general, the same trends
  - in some specific cases those trends are opposite
- these differences can be explained by the ideological differences between both newspapers
- LDA is a suitable tool even with degraded text, using the page as analysis unit instead the single news

Thanks!

Carlos G. Figuerola, University of Salamanca

figue@usal.es