European Language Resource Coordination
Connecting Europe Facility

Lingue e Tecnologie della Lingua in Italia

Simonetta Montemagni

Istituto di Linguistica Computazionale "A. Zampolli" - CNR

ELRC Technological Italian National Anchor

European Language Resource Coordination
Connecting Europe Facility

Le **lingue** in Italia

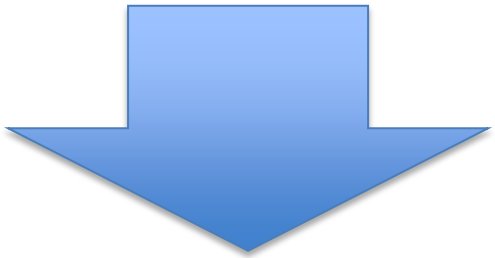
Parte 1

Il multilinguismo in Italia
Le lingue ufficiali

Legge 482/99 "Norme in materia di tutela delle minoranze linguistiche storiche"

Art. 1.
1. La lingua ufficiale della Repubblica è l'italiano. […]

Art. 2.
1. In attuazione dell'articolo 6 della Costituzione e in armonia con i principi generali stabiliti dagli organismi europei e internazionali, la Repubblica tutela la lingua e la cultura delle popolazioni albanesi, catalane, germaniche, greche, slovene e croate e di quelle parlanti il francese, il franco-provenzale, il friulano, il ladino, l'occitano e il sardo.



italiano [*occitanico*, *francese*, ***franco-provenzale***, *tedesco*, *sloveno*, ***ladino***, ***friulano***, *serbocroato*, *albanese*, *neogreco*, *catalano*, ***sardo***]

Il multilinguismo in Italia
Oltre le lingue ufficiali

- L'Italia ha una lingua ufficiale e 12 co-ufficiali

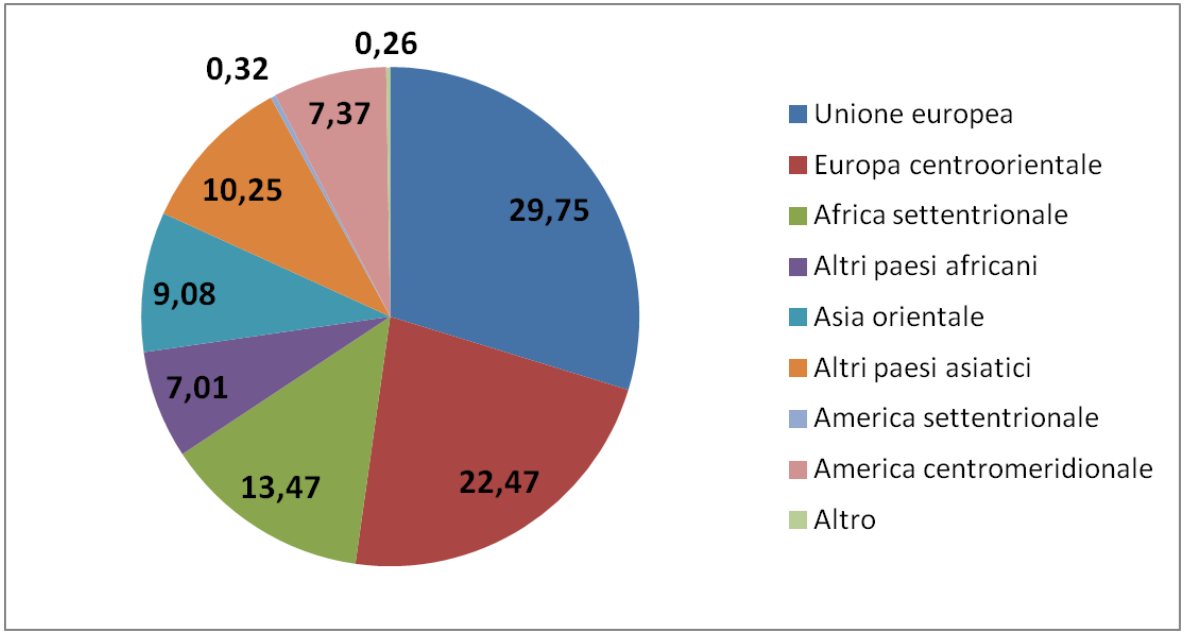
- La vita reale è un'altra cosa
  - Immigrazione
  - Commercio
  - Turismo
  - Lo sviluppo e la cooperazione tra le regioni d'Europa
    - Mobilità e traffico
    - Energia e cambiamento climatico
    - Ambiente e risorse naturali
  - Rapporti economici e legali transfrontalieri

Il multilinguismo in Italia
Le lingue immigrate

- Alle lingue ufficiali e co-ufficiali nella comunicazione tra cittadini e istituzioni si aggiungono le **lingue delle minoranze di nuovo insediamento**

- 2015: cittadini stranieri residenti in Italia 8,2% (+1,9% rispetto al 2014)
  - **Romania** 22,6%
  - **Albania** 9,8%
  - Marocco 9,0%
  - Cina Rep. Popolare 5,3%
  - Ucraina 4,5%
  - Filippine 3,4%
  - India 2,9%
  - Moldova 2,9%

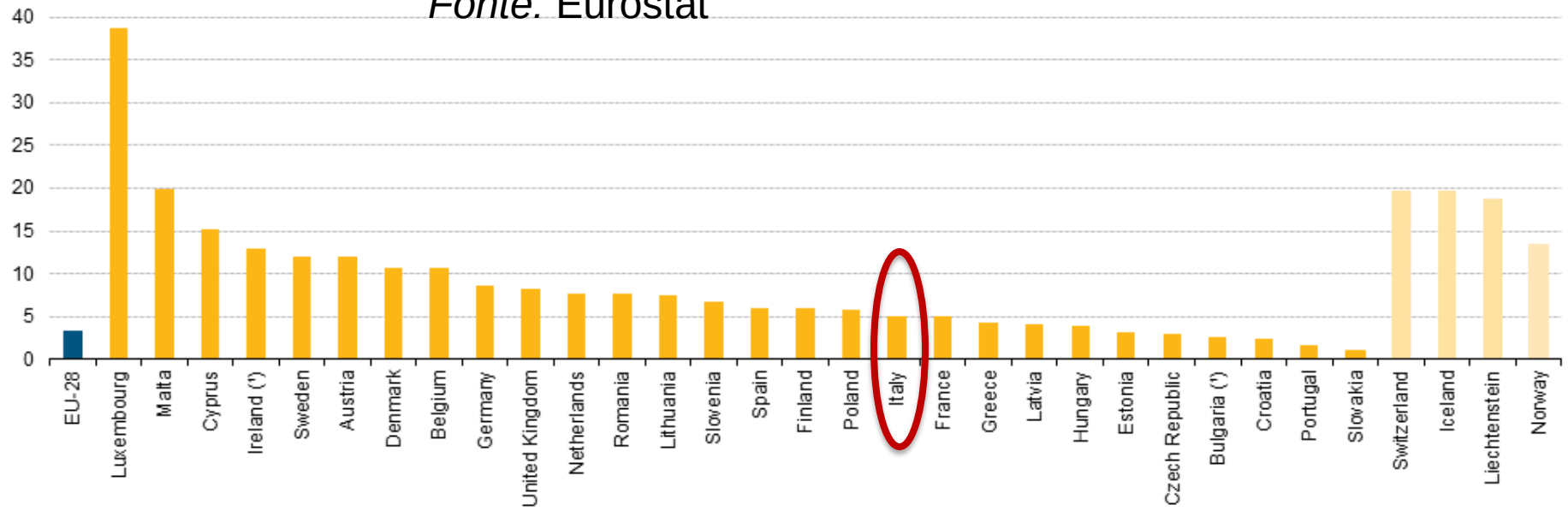
2015: Cittadini stranieri in Italia per area d'origine



Immigrazione: Italia vs Europa

Immigrati in Europa, 2013 (per 1 000 abitanti)

Fonte: Eurostat



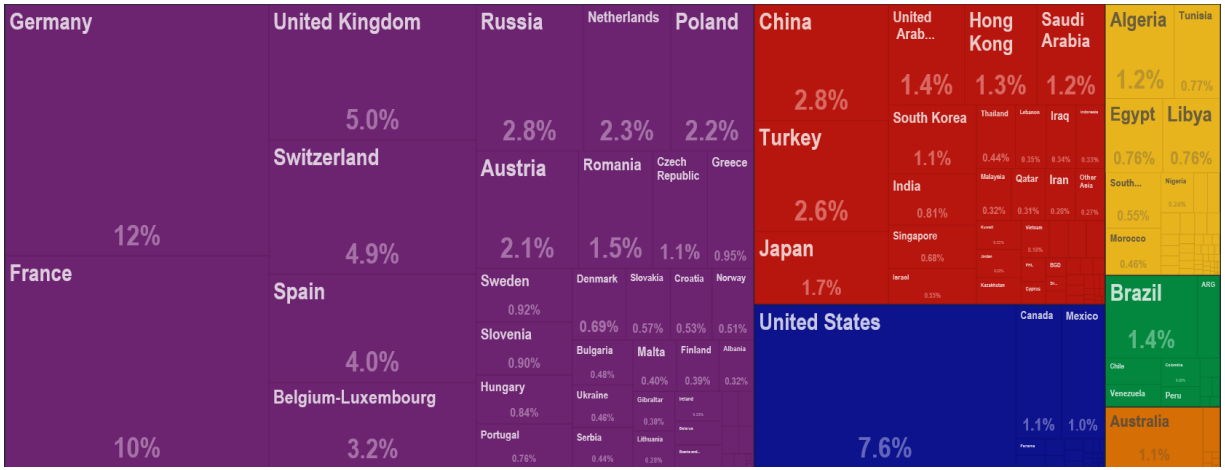
(') Provisional.
Source: Eurostat (online data codes: migr_imm1ctz and migr_pop1ctz)

European Language Resource Coordination
Connecting Europe Facility

Le lingue dei paesi confinanti

European Language Resource Coordination
Connecting Europe Facility

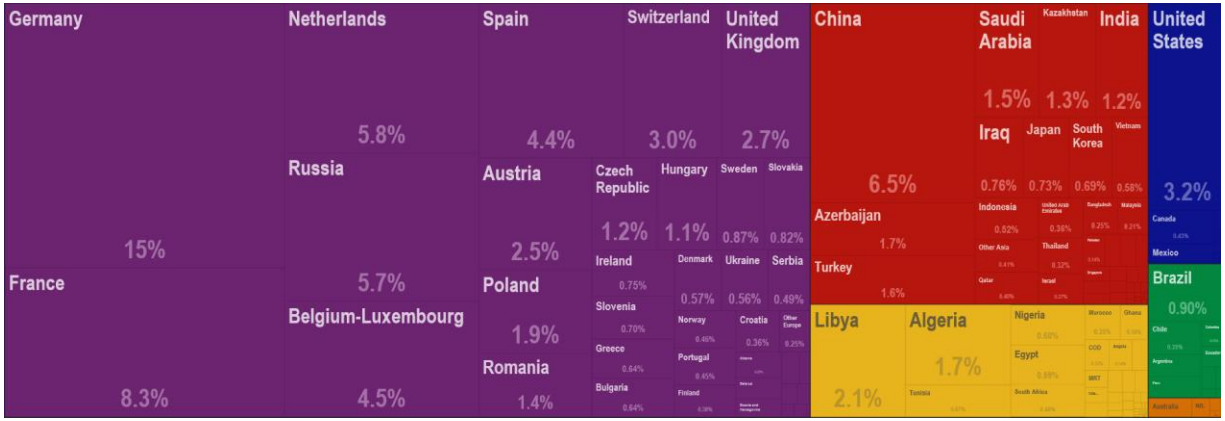
Relazioni commerciali dell'Italia (2013)



Esportazioni

Top 5

- **Germania (12%)**
- **Francia (10%)**
- USA (7,6%)
- UK (5%)
- Svizzera (4,9%)



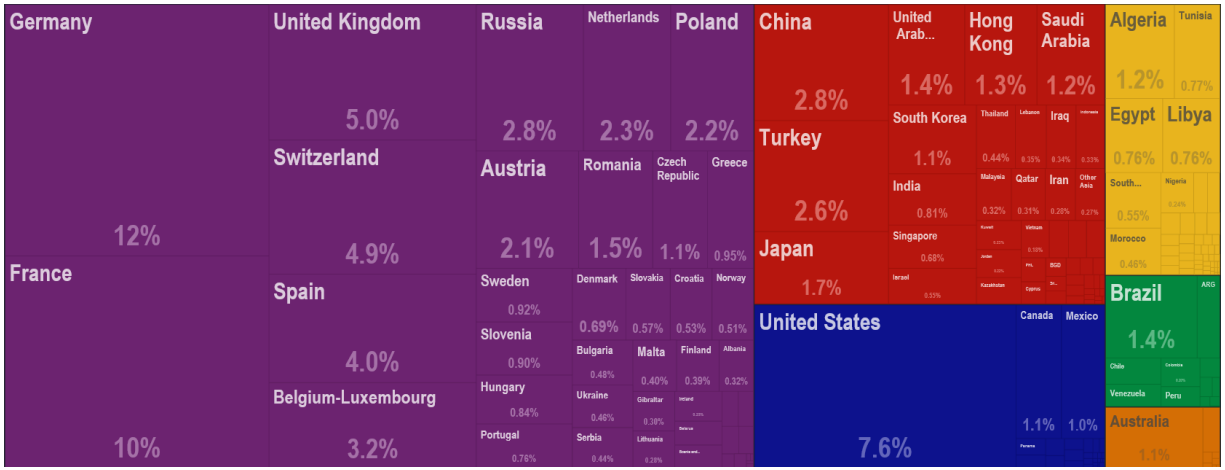
Importazioni

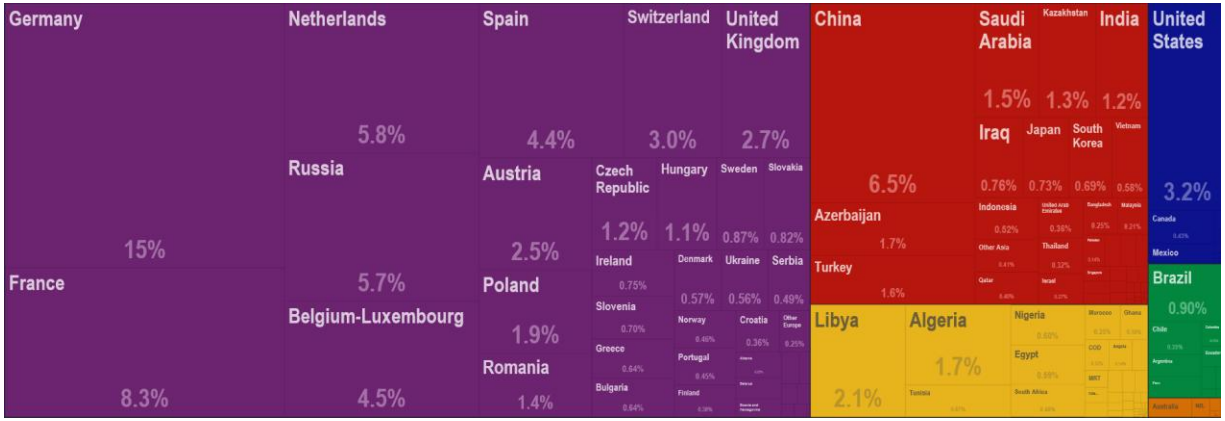
Top 5

- **Germania (15%)**
- **Francia (8,3%)**
- Cina (6,5%)
- Paesi Bassi (5,8%)
- Russia (5,7%)

Fonte: Observatory of Economic Activity

Relazioni commerciali dell'Italia (2013)



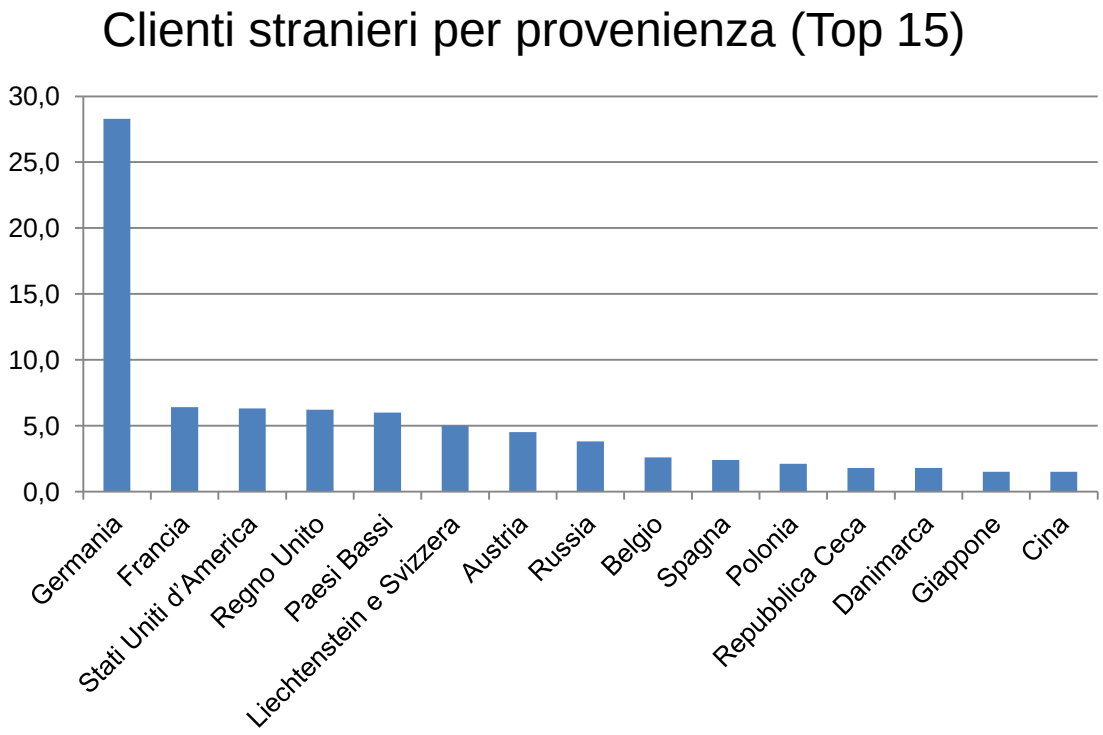


Principali lingue coinvolte

Tedesco
Francese
Inglese
Cinese
Olandese
Russo
…
Spagnolo
Arabo
Turco
Polacco

Fonte: Observatory of Economic Activity

Flussi turistici in Italia per paese di residenza (2013)



Fonte: Annuario Istat 2015

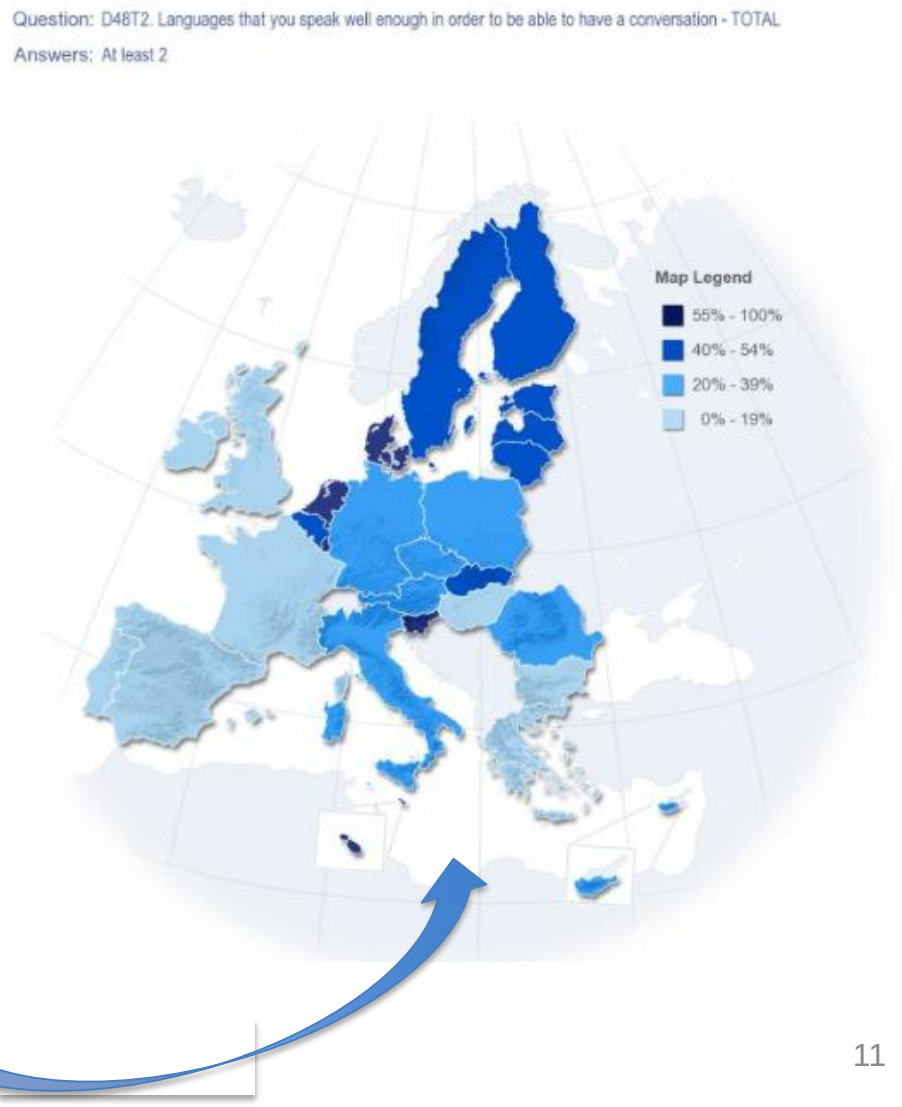
Multilinguismo vs plurilinguismo in Italia

Multilinguismo

- 1+12 lingue (co-)ufficiali
- Lingue immigrate
- Lingue coinvolte nelle relazioni di cooperazione e sviluppo internazionali, commercio e turismo

Plurilinguismo

- Fonte: Eurobarometer 386 “Europeans and their languages” (2012)
- Italiano: la seconda L1 in Europa (13%)
- Le tre L2 più parlate dagli italiani: Inglese, Francese, Spagnolo
  - 38% dichiara di parlare una L2
  - 62% dichiara di parlare **nessuna** L2
    - Italia penultima tra tutti i paesi UE, seguita dell’Ungheria (65%)
    - Media UE; 46%



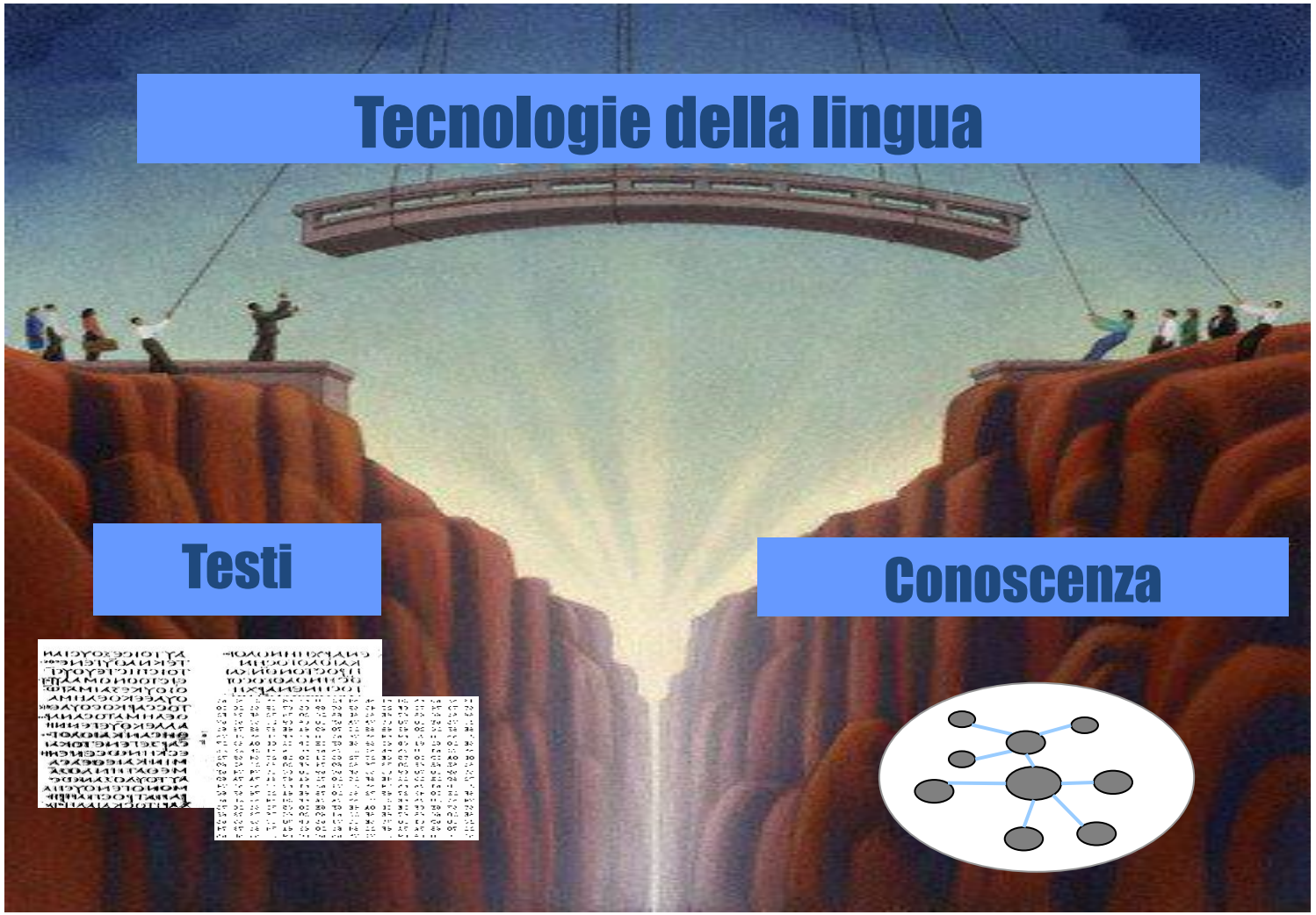
European Language Resource Coordination
Connecting Europe Facility

Le **tecnologie della lingua** in Italia

Parte 2

European Language Resource Coordination
Connecting Europe Facility

Tecnologie della lingua come “ponte” …



Le “insidie” del linguaggio per il computer

Nome o verbo?

La vecchia **porta** la sbarra

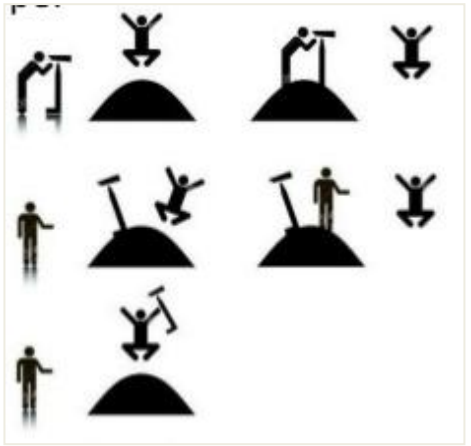
Quale senso di ***interesse***?

Il tasso di **interesse** è variabile anche in funzione della moneta di riferimento

Ha mostrato molto **interesse** per la Linguistica Computazionale

Ho visto l’uomo **sulla collina** ***con il telescopio***

Chi è sulla collina?
Chi ha il telescopio?



Le “insidie” del linguaggio per il computer: peculiarità dell’italiano

- Ordine libero dei costituenti della frase
  - *La lettera che ha scritto la segretaria*
  - *La segretaria che ha scritto la lettera*
- Omissione del soggetto
  - *Legge un libro* vs *Legge Maria*
- Italiano della Pubblica Amministrazione
  - Costruzione sintattica molto elaborata
  - Tendenza alla “frase unica”
  - Preferito l’uso di lessico astratto e arcaico

Italo Calvino, *Per ora sommersi dall’antilingua*, Il Giorno, 8 febbraio, 1965

Il brigadiere è davanti alla macchina da scrivere. L’interrogato, seduto davanti a lui, risponde alle domande un po’ balbettando, ma attento a dire tutto quel che ha da dire nel modo piú preciso e senza una parola di troppo: «Stamattina presto andavo in cantina ad accendere la stufa e ho trovato tutti quei fiaschi di vino dietro la cassa del carbone. Ne ho preso uno per bermelo a cena. Non ne sapevo niente che la bottiglieria di sopra era stata scassinata». Impassibile, il brigadiere batte veloce sui tasti la sua fedele trascrizione: «Il sottoscritto essendosi recato nelle prime ore antimeridiane nei locali dello scantinato per eseguire l’avviamento dell’impianto termico, dichiara d’essere casualmente incorso nel rinvenimento di un quantitativo di prodotti vinicoli, situati in posizione retrostante al recipiente adibito al contenimento del combustibile, e di aver effettuato l’asportazione di uno dei detti articoli nell’intento di consumarlo durante il pasto pomeridiano, non essendo a conoscenza dell’avvenuta effrazione dell’esercizio soprastante».

Identikit della lingua della PA
(D. Brunato, 2015)

Le parole	La struttura delle frasi	Il testo
• Pseudo-tecnicismi (o tecnicismi collaterali) (es. *balneazione, fattispecie*), • Termini arcaici (es. *allorchè, testè, suddetto, Signoria Vostra*) • Latinismi (es. *una tantum*, *pro capite*) • Forestierismi (es. *governance, back office, front office*) • Termini specifici e formali (es. *diniego* invece di *rifiuto*)	Stile nominale - Verbo semanticamente vuoto + nominalizzazione (es. *apporre la firma*, *sottoporre a controllo* anzichè il verbo semplice *firmare*, *controllare*); - Perifrasi verbali (es. provvedere a riscuotere) - Verbi di modo indefinito (gerundio, infinito e participio)	Autoreferenzialità (comunicazione orientata allo scrivente e non al lettore)
Abbreviazioni e acronimi	Enclisi pronominale con verbo finito (es. *dicesi*, *trattasi*, *vedasi*.)	Organizzazione testuale rigida, tipica della prosa legislativa
Nomi astratti con suffissi in –zione/-mento (es. *stipulazione, espletamento*), nomi deverbali, spesso a suffisso zero (es. *subentro, scorporo, utilizzo*), verbi denominali (es. disdettare)	Ordine non canonico degli elementi (es. aggettivo prima del nome, *suddetto modulo*) o forme di focalizzazione meno attestate nella lingua standard (es. *tali disposizioni riceveranno le amministrazioni..*)	Uso massiccio di legami anaforici e cataforici che rimandano ad altri elementi nel testo (es. *visto, considerato, suddetto, sottoindicato, in calce*)
Locuzioni preposizionali e/o congiuntivali complesse (es. *al fine di*, *a condizione che*)	Costruzioni grammaticali infrequenti (Frasi impersonali, al passivo, incisi e parentetiche, doppia negazione)	Interstestualità marcata (continuo riferimento a fonti esterne all'interno del documento)
Formule stereotipate e pleonastiche (es. *entro e non oltre*, *in riferimento all'oggetto*)	Predominanza dell'ipotassi sulla paratassi, con lunghe catene di frasi subordinate	Mancanza di coesione tra sezioni e paragrafi del documento.

Tecnologie della lingua
Come funzionano

- Paradigma dominante oggi rappresentato da sistemi basati su algoritmi di apprendimento automatico
  - algoritmo
  - **dati e risorse linguistiche**

Tecnologie della lingua
Come funzionano

- Paradigma dominante oggi rappresentato da sistemi basati su algoritmi di apprendimento automatico
  - algoritmo
  - **dati e risorse linguistiche**

- Per trattare nuovi domini e/o varietà d'uso della lingua
  - è necessario integrare l'evidenza su cui si basa il sistema con dati e risorse relativi al nuovo dominio / varietà di lingua

Tecnologie della lingua in Italia

- Come possono essere di aiuto nell'erogazione di servizi multilingui
  - **Traduzione**
  - Estrazione di conoscenza
  - Costruzione di risorse lessicali, terminologiche e ontologiche di dominio, memorie di traduzione
  - Classificazione documentale, per contenuto o genere testuale
  - Valutazione della leggibilità e supporto alla semplificazione del testo

- Tecnologie sviluppate in linea con lo stato dell'arte a livello internazionale
  - alta reputazione e visibilità internazionale della ricerca italiana

Tecnologie della lingua in Italia

META-NET

META-NET

- Rete di Eccellenza europea
  - 60 centri di ricerca in 34 paesi
  - Prima *Strategic Research Agenda for Multilingual Europe 2020 - SRA* (2013)

2015: Strategic Agenda for the Multilingual Digital Single Market, definita a partire da documenti strategici e piani di azione preparati nell'ambito di numerosi progetti europei, inclusa la META-NET SRA

Tecnologie per l'italiano vs altre lingue europee

	excellent	good	moderate	fragmentary	weak or no support
MT		English	French, Spanish	Catalan, Dutch, German, Hungarian, **Italian**, Polish, Romanian	Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish

Text Analysis	excellent	good	moderate	fragmentary	weak or no support
		English	Dutch, French, German, **Italian**, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian

Speech	excellent	good	moderate	fragmentary	weak or no support
		English	Czech, Dutch, Finnish, French, German, **Italian**, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian

Resources	excellent	good	moderate	fragmentary	weak/no support
		English	Czech, Dutch, French, German, Hungarian, **Italian**, Polish, Spanish, Swedish	Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene	Icelandic, Irish, Latvian, Lithuanian, Maltese

Gli attori: la comunità italiana delle tecnologie della lingua

- I gruppi di ricerca che si occupano di tecnologie del linguaggio in Italia sono numerosi, si estendono su tutto il territorio nazionale e operano sia nell'area umanistica che in quella informatica
  - **Istituto di Linguistica Computazionale "A. Zampolli" – CNR**
  - **Università di Pisa**
  - **Fondazione Bruno Kessler (FBK), Trento**
  - **Università di Roma Tor Vergata**
  - Università di Torino
  - Università di Trento
  - Università di Bolzano
  - Università di Venezia
  - Università di Bologna
  - Università di Napoli
  - Università di Bari
  - ...
- Ai quali si affiancano numerose ditte tra le quali
  - Expert System, CELI, Euregio Srl, Reveal Srl, Almawave, Synthema, consorzio italiano Semantic Valley, IBM, META srl

Gli attori: la comunità italiana delle tecnologie della lingua

ForumTAL: istituito all'inizio degli anni 2000, prima rete italiana che ha riunito diversi tipi di attori interessati al trattamento automatico della lingua

Associazione Italiana di Linguistica Computazionale: fondata nel 2015, riunisce la comunità italiana che opera nel settore sia in ambito accademico che industriale

- tra le sue finalità
  - promuovere attività scientifiche e formative nel campo della Linguistica Computazionale e delle applicazioni delle tecnologie del linguaggio, in particolare quelle **rivolte alla lingua italiana**
  - stabilire e consolidare i **legami con altre iniziative, nazionali, europee ed internazionali**
  - **promuovere la Linguistica Computazionale nell'ambito della politica nazionale**
- la rete dei soci distribuita su tutto il territorio italiano ma non solo
  - Università, CNR, Centri di Ricerca, aziende
  - All'interno di diversi settori disciplinari (linguistica, informatica, ingegneria)

Verso servizi pubblici senza barriere linguistiche …

- La comunicazione multilingue e lo scambio di documenti tra le pubbliche amministrazioni nazionali in Europa richiede la specializzazione del sistema di traduzione automatica CEF.AT in relazione ai domini e generi testuali da trattare

- I dati testuali e linguistici disponibili presso le pubbliche amministrazioni possono contribuire in modo significativo alla specializzazione dei risultati di CEF.AT


grazie
thank you
merci
arigatô
gracias
danke
shukran
благодаря
obrigado
hvala
tänan
dankie
takk
tak
tack
dank u
faleminderit
kiitos
xièxiè
paldies
diolch
dhanyavād
köszönöm
terima kasih
mahalo
grazzi
ďakujem
dziękuję
děkuji
ačiū
mh'gōi
баярлалаа
mulțumesc
gràcies
дякую
хвала
спасибо
choukrane
ngiyabonga
kop khun
teşekkür ederim
nandri
a dank