NORMALIZZAZIONE DEL CORPUS ITALIANO

DOCUMENTAZIONE E MANUALE PER I COLLABORATORI

04.06.2014

rivisto il 02.03.2015

SOMMARIO

1 TOKENS

Quando trattiamo in modo informatico gli SMS non parliamo più di parole ma di *tokens*. Questi sono generati in modo automatico dal software impiegato. Il software riconosce gli elementi separati da spazi, apostrofi o trattini, così come la maggior parte delle emoticon e dei segni di punteggiatura. Un *token* può quindi essere una sequenza di lettere alfabetiche, ma può anche costituirsi di cifre o altri segni o di una miscela di questi.

Quindi quando parliamo di normalizzazione deve essere ben chiaro il fatto che non normalizziamo delle parole ma dei *tokens*. È possibile che il numero di *tokens* dell'SMS non corrisponda al numero di *tokens* a livello normalizzato. Inoltre, è possibile che l'identificazione automatica dei *tokens* da parte del sistema non avvenga sempre in modo corretto. Questo tipo di errore deve essere corretto manualmente.

1.1 UN TOKEN NELL'SMS → PIÙ TOKENS NORMALIZZATI

Negli SMS, un *token* rilevato in modo automatico può in realtà corrispondere a due o più *tokens* a livello normalizzato. È ad esempio il caso degli acronimi come *tvb* (= ti | voglio | bene) oppure di forme fuse del tipo *giaccavento* (= giacca | a | vento) o *vabbè* (= va | be'). Questa modifica può venir effettuata facilmente durante la normalizzazione: nella casella che si trova direttamente sotto il *token* dell'SMS introduciamo le due o più parole che corrispondono alla norma, separandole con uno spazio.

1.2 PIÙ TOKENS NELL'SMS → UN TOKEN NORMALIZZATO

Quando a livello dell'SMS ci sono più *tokens* (ad esempio [*sta*] [*sera*] oppure [*oltr'*] [*alpe*]) che corrispondono ad un solo *token* a livello normalizzato, quest'ultimo viene introdotto nella prima casella al di sotto dei *tokens* dell'SMS ([*stasera*] [ ] oppure [*oltralpe*] [ ]) mentre le altre caselle rimangono vuote.

1.3 TOKENIZZAZIONE ERRONEA A LIVELLO SMS

Quando al livello SMS c'è un errore di tokenizzazione (ad esempio [*c*] ['] in due *tokens*, oppure nel caso di un'emoticon tokenizzato in modo errato [<] [3]) bisogna notarlo. Per fare ciò, dobbiamo scegliere nel riquadro "SMS Details" della software normalizzazione lo stato "needs retokenizing". Questo errore sarà corretto a mano più tardi. Quando utilizzeremo l'opzione "needs retokenizing", lasceremo vuote le caselle sottostanti gli elementi in questione.

2 MAIUSCOLE/MINUSCOLE

Tutte le parole sono scritte come si trovano in un vocabolario italiano (Lo Zingarelli online è stato utilizzato come vocabolario di riferimento). Concretamente significa che i nomi propri (inclusi gli acronimi), così come l'item *Dio* si scrivono con l'iniziale maiuscola. Delle entità politiche sono trattate come dei nomi propri in italiano: *la Repubblica italiana*, *il Gran Consiglio* (in caso di dubbio consultare il vocabolario di riferimento). Tutte le altre parole si scrivono in minuscolo. Questo vale anche per le parole che si situano ad inizio frase. L'unica eccezione è rappresentata dai sostantivi tedeschi che mantengono l'iniziale maiuscola.

3 APOSTROFO

L'apostrofo non è mai computato come un *token* a sé stante. Esso, trattandosi di un'elisione, deve sempre essere legato al *token* eliso. Una tokenizzazione corretta dell'apostrofo è rappresentata dall'esempio seguente [*com'*] [*è*] e quindi non da [*com*] ['] [*è*]. In quest'ultimo caso bisogna segnalare l'errore indicando "needs retokenizing" nel riquadro "SMS Details".

4 ELLISSI

Qualsiasi tipo di omissione che si trova nell'SMS deve anche figurare nel livello normalizzato. Non viene aggiunto nulla.

5 VARIANTI GRAFICHE

In linea di massima le varianti grafiche sono normalizzate.

Esempi:

(1) [ke] → [che]

(2) [ti] [kiamo] → [ti] [chiamo]

Quando il suono reso dalla grafia corrisponde alla pronuncia reale della parola (cfr. esempi (1) e (2)), non annoteremo questo *token* come abbreviazione. Al contrario verranno annotate come abbreviazioni le grafie che corrispondono alla pronuncia della lettera quando si cita l'alfabeto:

(3) [se] [**c**] [riesco] → [se] [ci] [riesco]

(4) [cerco] [**d**] [ricordarmi] → [cerco] [di] [ricordarmi]

(5) [semmai] [**t**] [spiego] → [semmai] [ti] [spiego]

Negli esempi (3), (4) e (5) i *token* [c], [d] e [t] verranno quindi normalizzati e marcati con l'attributo "Abbreviation".

Le forme del tipo *nexuna*, *proxima* e *adexo* per *nessuna*, *prossima*, e *adesso* vengono anch'esse marcate come abbreviazioni in quanto il suono reso dalla grafia non corrisponde alla pronuncia reale della parola.

6 PAROLE COMPOSTE

Le parole composte che sono state suddivise in diversi *tokens* sono ricostituite scrivendole per intero nella prima casella sotto i *tokens* in questione e lasciando vuote le successive:

(7) [niente] [popò] [di] [meno] → [nientepopodimeno] [ ] [ ] [ ]

7 FORME ABBREVIATE

7.1 PUNTI NELL'ABBREVIAZIONE

Quando un'abbreviazione contiene dei punti (come in *ecc.*, *cfr.*), essi devono essere uniti all'abbreviazione. Ciò significa che è necessario segnalare ogni errore indicando "needs retokenizing" nel riquadro "SMS Details":

(8) [ecc] [.] → [ecc.]

L'unica eccezione è quando il punto si trova alla fine di una frase. In questo caso non viene unito all'abbreviazione.

7.2 ABBREVIAZIONI FRASEOLOGICHE

Negli SMS le abbreviazioni sono utilizzate frequentemente. In generale si tratta di abbreviazioni largamente diffuse come *cmq* per *comunque* o *prox* per prossima, ma si possono trovare anche delle abbreviazioni che sono riconosciute unicamente dall'autore di un messaggio e dal suo interlocutore.

Nel nostro corpus l'obiettivo è di annotare tutte le abbreviazioni che ritroviamo in modo da poterle ritrovare facilmente in un secondo momento. Per questo motivo ogni *token* che contiene una forma abbreviata riceve l'attributo "Abbreviation". Gli acronimi (come *tvb* per *ti voglio bene*) e le abbreviazioni (*nn* per *non*) sono categorizzate allo stesso modo per delle ragioni di semplicità. Se un ricercatore si interessa alle forme abbreviate, può ritrovarle facilmente e di seguito classificarle secondo le proprie necessità.

Le forme abbreviate non standard (cioè non presenti nel vocabolario di riferimento) che possono essere indovinate facilmente e senz'alcun dubbio vengono risolte e riscritte nella loro forma completa al livello normalizzato. Tutte le forme abbreviate che vengono riscritte per intero ricevono comunque l'attributo "Abbreviation".

7.3 NOMI PROPRI ABBREVIATI, ACRONIMI, FORME ABBREVIATE LESSICALIZZATE

A lato delle forme abbreviate menzionate qui sopra, esistono anche altre forme abbreviate. È il caso di *OLAT* per *Online Learning And Training*, di *Laser* per *Light Amplification by Stimulated Emission of Radiation*, ma anche di *TI* per *Ticino* e di *dom* per *domenica* o *domani*.

Per decidere quale forma abbreviata verrà risolta e scritta per intero ci siamo posti le seguenti domande:

1. Che cosa è ragionevole per i collaboratori del progetto?

2. Cosa potrebbe risultare interessante per delle ricerche future?

Da qui risultano le regole seguenti:

a. Le forme abbreviate che si trovano come entrate principali nel vocabolario di riferimento non vengono risolte e scritte per intero (*Laser* resta *Laser*, *ecc.* resta *ecc.*, *min* resta *min*), ma ricevono comunque l'attributo "Abbreviation".
b. I nomi propri (salvo i toponimi) non sono analizzati (*ETH resta ETH*). Questo vale anche per i nomi dei club: *HCAP* (Hockey Club Ambrì Piotta) resta *HCAP*.
c. Le altre forme abbreviate che non presentano particolari problemi d'interpretazione sono scritte per intero (*TI* diventa *Ticino, lune* diventa *lunedì*, eccetera) e marcate come abbreviazioni.
d. Quando c'è un dubbio riguardo all'interpretazione (ad esempio quando compare la forma *dom* è spesso difficile per il ricercatore stabilire se si tratta della forma abbreviata per *domenica* o per *domani*) oppure quando la forma abbreviata non può essere analizzata,

essa è ripresa anche a livello normalizzato così come si presenta al livello SMS e verrà annotata con gli attributi "Abbreviation" e "Unclear".

e. I nomi di persona abbreviati (ad esempio *buona serata M)* non ricevono l'attributo "Abbreviation" perché non sono particolarmente rilevanti per la ricerca.
f. Le abbreviazioni che consistono in segni saranno marcate come abbreviazioni e scritte per intero:

(9) *x* → *per* ("Abbreviation")

(10) *+* → *più* ("Abbreviation")

Esempi:

Se in un SMS c'è la forma abbreviata *ecc* la indicheremo come "Abbreviation" e, dato che *ecc.* esiste come entrata nel vocabolario, adatteremo l'abbreviazione alla forma standard che ritroviamo nel vocabolario. La forma normalizzata sarà quindi *ecc.*

(11) *ecc* → *ecc.*

Nel seguente caso invece sceglieremo l'attributo "Abbreviation" e scriveremo l'abbreviazione per intero:

(12) *se trovo* ***qlcn*** *ti aspetto* → *se trovo* ***qualcuno*** *ti aspetto*

Per la forma abbreviata

(13) *OMG* → *oh my God*

Bisogna aggiungere che si tratta di un'abbreviazione inglese. Annotiamo quindi dapprima l'attributo "english" ed in seguito risolviamo la forma per intero *oh my God* grazie alla funzione "edit gloss". Infine non dobbiamo dimenticare di marcarla come abbreviazione.

7.4 SCRITTURE REBUS

Anche le scritture rebus – cioè quando segni e/o numeri sono utilizzati come equivalenti fonologici di una o più lettere – sono da considerare delle forme abbreviate.

(14) *dove* ***6****?* → *dove* ***sei****?*

(15) ***xke*** → *perché*

(16) ***xò*** → *però*

Questo tipo di scritture sono normalizzate, scritte per intero e indicate come abbreviazioni.

8 AFERESI, APOCOPE E ELISIONE

Per quanto riguarda aferesi e apocope, di regola vengono normalizzate le forme che non risultano come lemma principale nel dizionario di riferimento (in caso di dubbio verificare). Le forme apocopate vengono dunque normalizzate come segue:

(17) *buon* → normalizzato secondo le regole classiche della grammatica italiana (cfr. il N.B. alla voce *buono* nello Zingarelli online)

(17a) *nessun* → normalizzato secondo le regole classiche della grammatica italiana (cfr. il N.B. alla voce *nessuno* nello Zingarelli online)

(18) *son* → *sono*

(19) *andar* → *andare*

(20) *gran* → *grande*

Fanno eccezione le forme *far*, *pur, bel, ben* e *quel* in quanto sono registrate come entrate principali nel dizionario di riferimento. Ad ogni modo le forme apocopate non sono considerate come delle abbreviazioni e di conseguenza **non** ricevono l'attributo "Abbreviation".

Le **aferesi** del tipo *'sto*, *sto* e tutte le loro declinazioni vengono normalizzate e scritte in forma completa. Malgrado il fatto che queste forme siano presenti come entrata principale nel vocabolario di riferimento, abbiamo deciso di normalizzarle. Le aferesi non vengono annotate come abbreviazioni ma soltanto normalizzate come segue:

(21) *sto* → *questo*

(22) *sta* → *questa*

…

Le elisioni, in particolare degli articoli, a livello normalizzato rimangono esattamente come si presentano nell'originale. Vale a dire che *l'* rimane *l'* anche a livello normalizzato. La cosa più importante a cui prestare attenzione è che l'apostrofo sia incluso nello stesso *token* della forma elisa:

(23) [*l'*] [*apostrofo*] e **non** [*l*] [*'*] [*apostrofo*]

Nel caso in cui si riscontrano dei casi di tokenizzazione errata è necessario segnalarlo selezionando "needs retokenizing" nel riquadro "SMS Details".

9 NOMI

Il corpus è già anonimizzato. I nomi propri presenti non dovrebbero quindi più poter essere attribuiti agli autori e ai loro conoscenti. I cognomi sono stati sostituiti con <*Lastname*> mentre i nomi sono stati sostituiti a rotazione. Ciononostante questo processo di anonimizzazione non è ancora completo. Da un lato perché certi nomi, essendo omografi ad altre parole funzionali, non possono venir anonimizzati in modo automatico (ad esempio *Felice*). Dall'altro lato, è possibile che certi nomi siano sfuggiti all'anonimizzazione automatica alla quale il corpus è stato sottoposto. È quindi necessario controllare manualmente se il processo di anonimizzazione è completo.

9.1 COME RICONOSCERE SE UN NOME È STATO ANONIMIZZATO?

I nomi anonimizzati sono marcati come tali. Infatti essi risultano già copiati nel livello normalizzato. Se a livello normalizzato compare già lo stesso nome che al livello SMS significa che tutto è corretto e il nome è già stato anonimizzato. Al contrario, sei il nome non risulta anche a livello normalizzato bisogna marcare il *token* con l'attributo "missing anonymisation". Questo vale anche per quei nome omografi con altre parole (come per *Felice*). Questi casi saranno poi trattati manualmente.

Quando assegniamo ad un *token* l'attributo "missing anonymisation", la casella corrispondente resterà vuota (per delle ragioni pratiche).

9.2 NOMI ABBREVIATI

I nomi di persona come ad esempio in *buona serata M* non ricevono l'attributo abbreviazioni (vedi capitolo 7.3 sulle forme abbreviate). Non riceveranno nemmeno l'attributo "Unclear".

9.3 NOMI DI PRODOTTI, MARCHE

I nomi di prodotti, come ad esempio *Ferrari*, non vengono anonimizzati e nemmeno annotati. Sono semplicemente normalizzati graficamente.

(24) *facebuk → Facebook*

9.4 NOMI PROPRI STRANIERI

I nomi propri di canzoni, gruppi, film, locali, marche, ecc. stranieri sono ripresi così come sono (o vengono tutt'al più normalizzati nella lingua in questione se si tratta di abbreviazioni, vedi cap. 7.3) e non ricevono l'attributo della lingua corrispondente, a patto che non esista un loro corrispettivo in italiano (vedi anche cap. 13.6)

(25a) *hai voglia di andare a vedere* ***New Moon****?*

(25b) *stasera* ***Chat Noir****?*

(25c) *sulla pagina di* ***Facebook***

Nel caso di titoli di film, libri, ecc. che invece hanno un corrispettivo italiano, il nome verrà ugualmente ripreso così com'è, ma riceverà l'attributo della lingua corrispondente: nell'esempio seguente, *Tartuffe* riceverà dunque l'attributo di lingua francese, dato che il pezzo teatrale in italiano s'intitola *Il Tartufo*.

(25d) *sono andata a vedere* ***Tartuffe*** *di Molière*

9.5 TOPONIMI

Nei toponimi, ogni parola corrisponde a un *token*. [*Lago*] [*Verbano*] costituisce due *tokens*. Sono ripresi tale e quale anche a livello normalizzato a meno che non ci siano delle varianti ortografiche.

(26) [*Lago*] [*Verbanoooo*] → [*Lago*] [*Verbano*]

I toponimi non vengono adattati all'italiano standard. Essi ricevono l'attributo di lingua corrispondente solamente se si tratta di un toponimo che conosce una traduzione italiana standard.

(27) *Zürich → Zürich*

In questo caso il toponimo rimane *Zürich* ma riceve l'attributo di lingua "German" in quanto esiste una variante italiana del nome della città (*Zurigo*).

10 ONOMATOPEE

Come nei fumetti, negli SMS vengono utilizzate delle forme onomatopeiche come *whoaa*, *grrrr*, *mmmmh*, ecc. Queste forme saranno annotate come onomatopeiche.

10.1 DELIMITAZIONI

In linguistica il termine "onomatopea" è molto ampio. Esso include anche delle parole come *sussurrare* o *cucu*. A noi non interessa annotare quest'ultimo tipo di onomatopee. I limiti fra onomatopee *strictu sensu* e parole lessicali come *ululare* da un lato e le interiezioni come *ah* dall'altro non sono ben

definiti; in questo senso, abbiamo scelto come linea guida la distinzione operata nella voce "onomatopea" del vocabolario Treccani (consultato nella sua versione online): alle "onomatopee" *strictu sensu* vengono infatti opposte forme "di origine onomatopeica", che hanno subito "un completo adattamento grammaticale, con l'aggiunta di desinenze e suffissi che le rendono elementi stabili (soprattutto sostantivi e verbi) del lessico". Tali forme, come *ululare*, non verranno quindi da noi considerate come onomatopee.

Saranno definiti quindi come "onomatopea" i seguenti casi:

- Dei *tokens* che non sono repertoriati nel vocabolario di riferimento.
- Dei *tokens* che hanno un carattere onomatopeico, e che quindi sono indicati nel vocabolario di riferimento come *vc. onomat.* o come *vc. espressiva* con l'eccezione, come detto sopra, di tutte le forme che il Treccani indica non come "onomatopee", ma forme "di origine onomatopeica" (quindi consideriamo come onomatopea una forma come *ah*, o *grrr*, ma non come *sussurrare* o *ululare*).

In questo modo, individueremo tutte le onomatopee *strictu sensu*, indipendentemente dalla loro appartenenza al vocabolario di riferimento o dalla variante grafica con cui sono state proposte, con il vantaggio di poter facilmente rintracciare l'insieme di tali forme nel corpus.

Processo concreto

Concretamente, i *tokens* con carattere onomatopeico vengono cercati nel vocabolario di riferimento. Se vi si trovano (come ad esempio *ah*, *grr*), sono ripresi così come si presentano nel messaggio, e ricevono comunque l'attributo "onomatopeia" (*ONO*). Se non vi si trovano (come ad esempio *ahhhh*, *ehehehehe*, ma anche *pfffff*), si marcheranno ugualmente come *ONO*, e si normalizzeranno riconducendole a forme presenti nel vocabolario:

(28a) *ahhhh* → *ah*

(28b) *ahahah* → *ah ah ah*

Ove ciò non fosse possibile, i termini verranno ripresi così come sono, e riceveranno anch'essi l'attributo *ONO*:

(28c) *argh* → *argh*

11 ORTOGRAFIA DI RIFERIMENTO

L'ortografia di riferimento è quella dello Zingarelli online. Per i lessemi che hanno più varianti normative decideremo per una sola variante:

- È il caso di *mezzora* e *mezz'ora* che sono entrambe presenti nel vocabolario di riferimento. In questo caso è stata scelta l'entrata principale, quindi *mezzora*.
- Per quanto riguarda *cd/CD e pc/PC* è stata scelta la variante minuscola in quanto la sigla maiuscola presentava più significati possibili. La forma scelta tra *sms/SMS* e *dvd/DVD* è quella maiuscola.

Attenzione: *SMS, cd* e *pc* sono anche da annotare con l'attributo "Abbreviation".

12 ERRORI DI BATTITURA E D'ORTOGRAFIA

Nel caso in cui si ritrova un errore di battitura o d'ortografia evidenti (ad esempio delle consonanti raddoppiate o semplici che non corrispondono all'ortografia dell'italiano standard) correggeremo l'ortografia.

(29a) *Gracie mille → Grazie mille*

13 ERRORI DOVUTI A "SCARSA COMPETENZA"

In caso di errori dovuti all'utilizzo involontario di forme estranee allo standard (dovuti quindi ad una "scarsa competenza" della lingua, ad una momentanea negligenza, o all'utilizzo di un registro particolarmente trascurato), agiremo come segue:

Se l'errore consiste nell'utilizzo inappropriato (secondo la norma dell'italiano) di una forma esistente (si tratta quindi della scelta di un vocabolo che esiste, ma non sarebbe secondo la norma da usare nella specifica frase in questione), allora il termine verrà ripreso così com'è e non riceverà nessun attributo:

(29b) *la nascita **da** Giovannino → la nascita **da** Giovannino*

Se invece l'errore consiste in una costruzione inesistente, i casi sono due: se è evidentemente riconoscibile una radice straniera alla quale è stata applicata la morfologia italiana (es: *lessiva*), allora si seguirà il procedimento esposto nel cap. 14.4; se invece la costruzione è del tutto errata o irriconoscibile, allora la si normalizzerà nella parola italiana che lo scrivente vuole esprimere:

(29c) ***feriamo → facciamo***

(Se al posto di *feriamo* avessimo trovato qualcosa come *faisiamo*, allora non sarebbe stato normalizzato, ma sarebbe stato mantenuto e marcato come *fra*, secondo la regola del cap. 14.4).

Si noti inoltre che nel caso di errori legati unicamente agli accordi morfologici, i termini verranno normalizzati secondo la corretta flessione italiana (vedi capp. 19-20).

14 LINGUE

Gli SMS sono già annotati per lingua, vale a dire che per ogni SMS è stato stabilito quale è la lingua principale e la lingua dei prestiti e dei prestiti ad-hoc (*nonce borrowings*) che si ritrovano nel messaggio stesso. Questa identificazione è stata effettuata a livello dell'SMS, mentre i *tokens* stessi non sono ancora stati annotati in base alla lingua. Questo compito è da fare in questa sede soltanto per quanto riguarda i prestiti **ad-hoc**. I *tokens* del tipo *kiss* o *happy* devono ricevere l'attributo di lingua inglese.

14.1 LINGUE CONSIDERATE

Durante l'annotazione della lingua dell'SMS diverse lingue sono state annotate. Partendo dall'idea che la ricerca nei nostri corpus verrà effettuata per le lingue nazionali, i dialetti e l'inglese principalmente, abbiamo deciso di limitarci a queste lingue. I prestiti ad-hoc di altre lingue saranno marcati con l'attributo di lingua *other*.

Gli pseudo-prestiti in pseudo-lingue creati intenzionalmente allo scopo di ottenere un effetto ludico, ironico, espressivo o simili, come ad esempio *fratellen* (che non è né la forma corretta italiana né quella tedesca) sono catalogati sotto *other*.

14.2 Procedimento concreto

Quando un *token* è annotato con un attributo di lingua, il sistema lo copia direttamente nel livello normalizzato. È quindi raccomandato, per prima cosa, di annotare l'attributo di lingua.

14.3 Adattamento dell'ortografia

Gli SMS presentano a volte un'ortografia vicina al parlato, cioè sono spesso scritti in un'ortografia pseudo-fonetica. Per questo ci capita di incontrare forme come *Tchuss* per *tschüss*. Quando possibile queste forme sono normalizzate secondo la lingua dalla quale provengono. A livello normalizzato scriveremo quindi *tschüss* secondo la forma riportata dal dizionario di riferimento utilizzato per il tedesco (Duden).

(30) *Tchuss → tschüss*

✋ Quando la lingua attribuita è un **dialetto** italiano o svizzero-tedesco, lasceremo la parola così come appare nell'SMS. Questo perché non esiste una norma nel dialetto e quindi risulta impossibile normalizzarlo.

14.4 Adattamento morfologico

Quando un elemento di una lingua straniera mostra una flessione italiana nell'SMS, questi tratti morfologici rimarranno invariati, ma si apporrà al nome l'attributo della lingua dalla quale proviene la radice del vocabolo. Se un autore scrive

(31a) *stas valigia e **lessiva** → stasera valigia e **lessiva***

(31b) *ciao **guapi**! → ciao **guapi**!*

lessiva e *guapi* verranno ripresi così come sono, e verrà loro assegnato rispettivamente l'attributo della lingua francese e spagnola. Se però la radice straniera è contemplata come prestito nel vocabolario di riferimento, non si metterà nessun attributo di lingue straniere:

(31c) *t'ho **mailato** la notizia*

dato che *mail* è presente nello Zingarelli, la parola non riceve alcun attributo.

Se non c'è nessuna flessione di partenza non verrà aggiunta nessuna flessione italiana. Se la flessione corrisponde alla forma prevista dalla norma della lingua di partenza, la lasciamo nella sua forma originale. Al contrario, se non corrisponde, verrà adattata alla norma di tale lingua:

(32) *non ci sono **problem** → non ci sono **problems***

14.5 Omofonia con parole italiane e tra lingue straniere

Quando c'è il dubbio a proposito dell'origine di una parola, come può essere il caso tra l'italiano e il dialetto, il contesto immediato è decisivo. Se la parola in questione si trova in una sequenza chiaramente attribuibile a una delle due lingue, la parola in dubbio è annotata come appartenente a quella lingua. Quando si tratta di parole isolate opteremo per l'italiano.

(33) *ricordòm che **stasera** devi domandat na roba* → isw (dialetto)

(33) ***stasera*** *ci sono anch'io* → ita

(34) ***stasera*** → ita

14.6 TOPONIMI E NOMI PROPRI

I toponimi in lingua straniera (ad esempio *New York*, *Zürich*, *Nyon)* vengono ripresi così come sono nel messaggio originale. Riceveranno l'attributo di lingua straniera soltanto se esiste un corrispondente in lingua italiana. Spesso risulta impossibile trovare delle delimitazione fra le lingue. *Berlin* ad esempio corrisponde al nome tedesco, a quello francese e a quello inglese della città, ma non a quello italiano *Berlino*. In casi come questo *Berlin* riceverà l'attributo di lingua tedesca perché è la lingua della città. *Nyon* e *New York* non avendo dei corrispettivi italiani **non** verranno annotati come prestiti ad-hoc.

Per quanto concerne i nomi propri in lingua straniera (*IBM*, *EPFL*), essi verranno ripresi così come si presentano nel messaggio originale e non verranno annotati come prestiti ad-hoc. *IBM* non viene dunque marcato come prestito in quanto non sembra particolarmente rilevante e interessante per eventuali ricerche (vedi anche cap. 9.4).

15 ORARI

In italiano a volte gli orari vengono rappresentati nella forma seguente: *21h00*. Questa forma viene riconosciuta come un singolo *token* e la lasceremo così com'è. Se ci troviamo di fronte ad una forma del tipo *21h* il *21* e la *h* risultano anch'essi come un singolo *token*. In questo contesto non verrà nemmeno marcato come abbreviazione.

16 "UNCLEAR"

L'attributo "unclear" è riservato alle forme per le quali è impossibile identificare che cosa l'autore voglia dire. Può trattarsi di un errore di battitura che rende la forma non interpretabile, di parole in lingue sconosciute o di forme omofone che non sono chiaramente interpretabili all'interno del sintagma (dunque di forme ambigue). Tutte queste forme saranno riprese così come appaiono nell'SMS e verranno marcate con l'attributo "unclear".

Esempio:

(35) *geo*

L'abbreviazione *geo* sarà marcata come "unclear" (e come abbreviazione) perché è impossibile stabilire se si tratta di *geografia*, *geologia*, *geometria*, ecc.

17 CIFRE E NUMERI

I numeri non vengono cambiati se appaiono come cifre nell'SMS

(36) *2 volte → 2 volte*

Se il numero è utilizzato come rebus (cfr. capitolo 7.4) allora le cifre vengono risolte e scritte in lettere.

I numeri ordinali abbreviati come 1°, 2°, 3°, ecc. sono scritti per intero e vengono indicati come abbreviazioni:

(37) *1o → primo*

18 AGGLUTINAZIONI

Al livello di normalizzazione, le parole sono separate come indicato nel vocabolario di riferimento.

18.1 AGGLUTINAZIONI CREATIVE

Quando gli elementi (che hanno una relazione sintagmatica) sono, non conformemente alla norma, separati da un trattino, manterremo questa ortografia per rendere conto della volontà dell'autore di marcare in modo esplicito questo passaggio come un unità:

(38) [*pulcino-canarino-mele-sulla-crostata*] → [*pulcino-canarino-mele-sulla-crostata*]

Se però ci sono delle varianti ortografiche nei sotto-segmenti, queste dovranno comunque venir normalizzate.

19 CONIUGAZIONE VERBALE

Durante la normalizzazione dei messaggi non vogliamo interpretare. Per questo motivo correggeremo la forma verbale rispettiva soltanto quando non c'è alcun dubbio per esempio riguardo ad un errore che si trova nell'accordo soggetto-verbo. Le correzioni si basano unicamente su forme chiaramente identificabili come false da un punto di vista grammaticale. Inoltre, non vogliamo perdere delle informazioni che potrebbero essere testimoni di un cambiamento in corso nella lingua, come ad esempio la sostituzione del congiuntivo con l'indicativo o la perdita di distinzione tra futuro e condizionale:

Da ciò quindi scaturiscono le seguenti regole:

a. I modi verbali "sbagliati" non vengono cambiati: l'indicativo rimane anche nelle subordinate che necessitano il congiuntivo, il condizionale non viene cambiato anche se intuitivamente ci si aspetterebbe un futuro, ecc.
b. Gli errori delle desinenze personali (persona e numero) vengono correte perché sono chiaramente riconoscibili, dal momento che c'è un soggetto nella stessa frase.
c. L'accordo di genere sul participio è corretto soltanto quando è chiaramente riconoscibile come non accordato secondo le regole della grammatica italiana.

20 ACCORDI

Gli accordi di numero e di genere per aggettivi, nomi, articoli e participi vengono corretti laddove risultano sbagliati.

(39) *baci pieno d'amore → baci pieni d'amore*

21 AGGETTIVI UTILIZZATI COME AVVERBI

Non modificheremo gli aggettivi che sono utilizzati come avverbi, anche se non presentano la forma avverbiale.

(40) *vado al cinema diretto → vado al cinema diretto*

In questo caso quindi *diretto* **non** viene cambiato in *direttamente*.