Term Variation in Institutional Languages: Degrees of Specialization in Municipal Waste Management Terminology

Cirillo Nicola, Vellutino Daniela
University of Salerno
84084 Fisciano, SA, Italy
{nicirillo, dvellutino}@unisa.it

Abstract

Institutional Italian is a variety of Italian used in the official communications of institutions, especially in public administrations. Besides legal and administrative languages, it comprises the language used in websites, social media, and advertising material produced by public administrations. We show that standard measures of lexical complexity alone, like the percentage of basic vocabulary, may be misleading when used for delineating the lexical profile of institutional languages and should be complemented with the examination of terminological variants. This study compares the terminology of three types of institutional texts: administrative acts, technical-operational texts, and informative texts. In particular, we collected 82 terms with various degrees of specialization and analysed their distribution within the subcorpora of ItaIst-DdAC_GRU, a corpus composed of institutional texts drafted by Italian municipalities about municipal waste management. Results suggest that administrative acts employ high-specialization terms compliant with the law, often in the form of acronyms. Conversely, informative texts contain more low-specialization terms, privileging single-word terms to remain self-contained. Finally, the terminology of technical-operational texts is characterised by standardized and formulaic phrases.

Keywords: institutional languages, terminological variation, text simplification

1. Introduction

Information and communication activities of the institutions have reshaped the sociolinguistic space of contemporary Italian. In recent years, a new variety of Italian language emerged: institutional Italian (Vellutino et al., 2012; Vellutino, 2018). In public administrations, this linguistic variety incorporates and redefines the historically attested variety of administrative bureaucratic Italian (Sobrero, 1993; Piemontese, 1999; Raso, 2005; Cortelazzo, 2021). Institutional Italian is used within the official communications of institutions in Italy and other countries that have Italian as their official language, e.g., the Swiss Confederation (Ferrari and Pecorari, 2022).

Vellutino et al. (2012); Vellutino (2018) represent the uses of institutional Italian, revisiting the model of sociolinguistic variation of contemporary Italian, originally proposed by Berruto (1987).

In public administrations, institutional Italian has different socio-pragmatic uses as displayed in Figure 1. They range from the specialized communication of the institutional languages of law and administration (i.e., special institutional languages) to institutional languages that use the media for conveying public and institutional information and communications (i.e., media institutional languages).

Vellutino et al. (2012); Vellutino (2018) proposed a classification model of institutional texts – CPI model (Comunicazione Pubblica e Informazione istituzionale ‘public communication and institutional information’) – which distinguishes the texts of the special institutional languages of law and administration from the texts of the institutional languages for information and communication, considering the different pragmatic-communicative contexts linked to the purposes of the discipline of public information and communication activities of administrations, defined by the relevant Italian law (Legge 150/2000; Legge 69/2009; D.lgs 33/2013; D.lgs 96/2017) and European Union regulations, in particular, about structural funds and recovery and resilience facility (EU Regulation 2021/1060, Next Generation Europe).

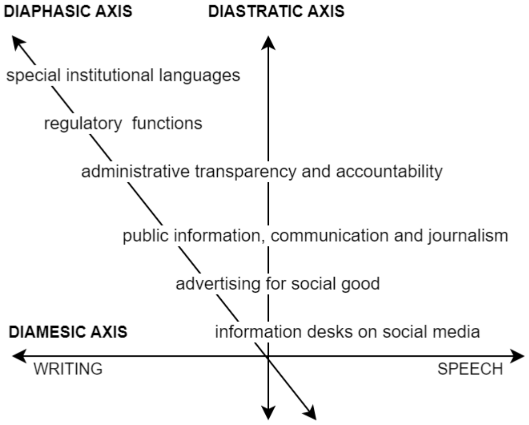
DIAPHASIC AXIS
DIASTRATIC AXIS
special institutional languages
regulatory functions
administrative transparency and accountability
public information, communication and journalism
advertising for social good
DIAMESIC AXIS
information desks on social media
WRITING
SPEECH

Figure 1: Socio-pragmatic uses of institutional Italian (Vellutino, 2018).

From a typological-structural point of view, the linguistic variety of institutional Italian is characterized by a rich textual repertoire and an endless neological dynamism due to the ongoing entry of specialized terminologies, often multiword expressions, which can also be reduced to acronyms, giving rise to lexical variants with different degrees of specialization (Serianni, 2007; Vellutino, 2018). High-specialization terms are known only to a close circle of specialists while low-specialization terms, being known to well-educated speakers, mix with the general lexicon and form a grey area between general and special languages (Gualdo and Telve, 2011).

An example of the mechanisms that form institutional terms involves the term *credito formativo* 'training credit'. This institutional term can further specialize for a specific domain of knowledge through an adjective: *credito formativo universitario* 'university training credit'. This second terminological formation can then be reduced to the acronym *CFU*, which is part of a jargon, from a sociolinguistic point of view.

In institutional texts, multi-word terms are phrases carrying a specific meaning. They can be considered a signal not only of the use of terminology but also of the transition from the "rigidity" of the text types of legal advertising, characterized by a special lexicon, to the"flexibility" of the text types of public and institutional information and communication.

This study aims to delineate the lexical profile of the text types defined in the CPI Model. Namely, we try to answer the following research questions.

- How complex is the lexicon of the different institutional text types?
- Is the percentage of basic vocabulary alone a good indicator of lexical complexity in institutional Italian?

To answer these questions, we examine the distribution of term variants, with different degrees of specialization, in a corpus of institutional texts about municipal waste management, produced by Italian municipalities

The remainder of this paper is organized as follows. Section 2 illustrates previous studies about administrative Italian. Section 3 outlines the methodology and the language resources used in the study. Within Section 4, we present and discuss experimental results. Section 5 provides conclusions.

2. Related Work

Administrative Italian has always been known for posing readability issues that hinder citizens from accessing public information. Nevertheless, despite the numerous simplification efforts, the problem is far from being solved (Lubello, 2018).

Attempting to improve the communication between public administrations and citizens, many authors provided essential guidelines addressing the simplification of administrative texts (Fioritto, 1997; Vellutino, 2018; Cortelazzo, 2021). Their key suggestions are the following:

- Use short sentences.
- Respect the subject-verb-object order.
- Avoid subordinate clauses, preferring coordination.
- Avoid the passive voice.
- Use common tenses.
- Use a basic vocabulary.
- Avoid technical terms when possible, otherwise, explain them.

The *Vocabolario di Base* 'basic vocabulary' VdB (De Mauro and Chiari, 2016) categorized Italian words based on their accessibility to speakers, defining three distinct classes: fundamental lexicon (approximately 2,000 lexemes); high-usage lexicon (approximately 3,000 lexemes); and high-availability lexicon (approximately 2,500 lexemes). The fundamental lexicon covers 86% of the word occurrences, the high-usage lexicon accounts for 6%, and the remaining 28,000 lexemes collectively contribute 8%.

From the perspective of natural language processing, various strategies and tools automatically measure text complexity and assign readability scores by analysing lexical and syntactic features. The GULPEASE index (Lucisano and Piemontese, 1988) exploits the length of words (in character) and the length of sentences (in words) to estimate the readability of a text. It also includes an interpretation scale, based on empirical tests. In addition, the Read-It tool (Dell'Orletta et al., 2011) combines statistical text features with lexical and syntactic information obtained from the VdB and the dependency graph of a sentence.

Corpora are another essential resource for the study of institutional languages. PAWaC (Passaro and Lenci, 2019) is a web corpus composed of administrative documents from the websites of Tuscan municipalities. SIMPITIKI (Tonelli et al., 2019) and Admin-It (Miliani et al., 2022) are parallel corpora containing original sentences and related simplified versions, obtained with various simplification strategies.

3. Material and Methods

3.1. ItaIst-DdAC_GRU corpus

The corpus employed in this study is ItaIst-DdAC_GRU (Vellutino and Cirillo, 2024), a corpus of administrative, technical and informative texts drafted by Italian municipalities about municipal waste management.

The texts were collected by the students of the course "Public Communication and Institutional Languages" at the University of Salerno. They collected the documents from the website of their municipality of residence or, when not available, requested them, exercising the right of simple civic access. Then, the documents were classified according to the CPI Model (Vellutino et al., 2012; Vellutino, 2018).

ItaIst-DdAC_GRU is divided into four subcorpora: *admin*, *tech*, *acc*, and *info*. Table 1 summarises the corpus composition. Being too small, the *acc* subcorpus has not been considered in this study.

The *admin* subcorpus is composed of administrative acts, mainly resolutions, forms and ordinances. The *tech* subcorpus includes technical-operational texts like MUD[1] and PEF[2] documents. The *info* subcorpus comprises informative texts like public notices, calendars and guides for the separate collection.

3.2. List of term variants

To select the terms for the analysis, we started from a list of words and phrases automatically extracted from the ItaIst-DdAC_GRU corpus through the Sketch Engine[3] keyword extraction tool (Kilgarriff, 2009). From this list, we selected only the terms with a consistent number of variants.

Moreover, the list was enriched by finding longer phrases derived from known terms with the aid of the collocation tool of Sketch Engine. E.g., from the term *centro comunale di raccolta 'municipal recycling centre'* we found its variant *centro comunale di raccolta <u>dei rifiuti</u>* 'municipal <u>waste</u> recycling centre'.

The final list contains 82 terms, expressing 6 concepts (see Appendix A).

3.3. Experimental tests

For the purpose of delineating the lexical profile of institutional languages, we conducted three experiments on the *admin*, *tech* and *info* subcorpora of ItaIst-DdAC_GRU.

3.3.1. Experiment 1

Experiment 1 aims to assess the complexity of the lexicon of each subcorpus, without considering terminology. In this experiment, lexicon complexity is modelled as the percentage of words from the basic vocabulary (VdB). The fewer VdB words a corpus contains, the more complex its lexicon. Moreover, the inner composition of the VdB also plays a role, high-availability words are more complex than high-usage words while fundamental words are the simplest. We also compared the percentage of VdB words with another index of lexical complexity: the Type-Token Ratio (TTR), which measures the richness of vocabulary.

The metrics mentioned above are computed via the Read-It tool[4] (Dell'Orletta et al., 2011). Being the full corpus too big to be processed by Read-It, this test was conducted on a simple random sample of 100 sentences from each subcorpus. In addition, we compared the results with a baseline extracted from the web corpus itTenTen20 (Jakubíček et al., 2013) by selecting 100 random sentences containing the article *il* 'the'.

3.3.2. Experiment 2

In experiment 2, we analysed the distribution of single-word terms, multi-word terms, and acronyms throughout the subcorpora.

Therefore, for each subcorpus, we calculated the relative frequency of the collected terms, grouping them by structure (i.e., single-word, multi-word, acronym). Moreover, we determined the significance of the observed association between term structures and text types through the chi-square test of independence.

3.3.3. Experiment 3

The goal of experiment 3 is to identify the features of the terminology used in each institutional text type

To this end, we computed the frequency of the collected terms in each subcorpus and, from the contingency table, we calculated the difference between observed and expected frequency.[5] Finally,

[1] *Modello Unico di Dichiarazione Ambientale* 'unified model for environmental declaration'

[2] *Piano Economico Finanziario* 'Economic and financial plan (of the separate collection service)'

[3] `https://www.sketchengine.eu/` accessed on 6 March 2024

[4] `https://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest` accessed on 6 March 2024.

[5] A positive value means that the term occurs in a subcorpus more times than expected under the null hypothesis (i.e., the hypothesis that a term is evenly distributed across the subcorpora). Conversely, a negative value means the term occurs fewer times than expected.

Subcorpus	Text type	Documents	Sentences	Tokens
admin	Administrative acts	140	24,193	1,021,131
tech	Technical-operational texts	26	4,279	183,773
acc	Texts for accountability	13	451	22,133
info	Informative texts	126	5,045	152,806
TOT		306	33,959	1,379,843

Table 1: Itaist-DdAC_GRU corpus.

we qualitatively analyzed, for each subcorpus, the most associated terms expressing a given concept.

4. Results ad Discussion

The results of experiment 1 are shown in Table 2. They seem to indicate that the lexicon of the *info* subcorpus is the most complex. It has a lower percentage of VdB than *admin*, the lowest percentage of fundamental lexicon and the highest percentage of high-availability lexicon.

Nevertheless, the TTR does not support this hypothesis. The *info* subcorpus has the lowest TTR, even lower than the baseline. The reason may be that a specialized corpus theoretically needs fewer lexemes than a web one since it is about a single topic. However, administrative acts and technical-operational texts compensate by employing a more sophisticated vocabulary, while the vocabulary of informative texts is relatively simple.

If we interpret the results of experiment 1 considering that terminology plays a significant role in specialized corpora, the high percentage of VdB in administrative acts may be attributed to their verbose nature. Conversely, informative and technical-operational texts contain fewer regular words and more terms, because they express concepts more concisely. Moreover, the fact that the *info* subcorpus contains many high-availability words suggests that informative texts use more low-specialization terms, some of which fall within the high-availability lexicon. From this perspective, informative texts have the simplest lexicon.

Figure 2 shows the results of experiment 2. There is a significant difference in the distribution of term structures throughout the text types (df=4, χ^2=520.33, p<0.001): single-word terms are preferred in informative texts; multi-word terms appear mostly in technical-operational texts and administrative acts; and acronyms are more frequent in administrative acts and informative texts.

4.1. Results of experiment 3

The concept <*centro di raccolta*> 'waste recycling centre', in administrative acts is mostly conveyed through the acronym *CRC* (+137) and the term *centro di raccolta* 'recycling centre' (+110), as defined in the Italian legislation. Widely used are also the acronym *CdR* (+74), and the variant *centro di raccolta comunale* 'municipal recycling centre' (+39). In contrast, informative texts are characterised by the more colloquial variants *isola ecologica* lit. 'ecological island' (+132) and *ecocentro* 'ecocentre' (+46). Technical-operational texts do not possess any strong relationship with any term expressing this concept.

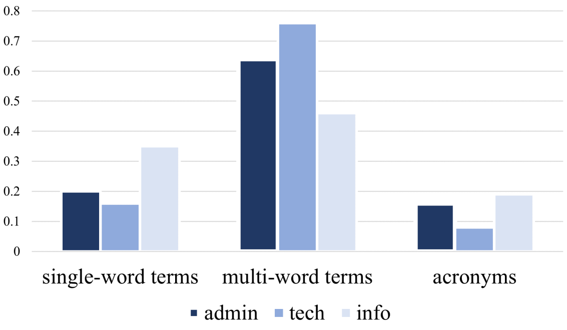

Figure 2: Distribution of term structures in the subcorpora of ItaIst-DdAC_GRU.

The concept <*rifiuto organico*> 'organic waste' is mostly conveyed in administrative acts through the term *frazione organica* 'organic fraction' (+31). In informative texts, the preferred variants are the singe-word terms *umido* 'wet waste' (+138) and *organico* 'organic waste' (+63). Technical-operational texts extensively use the term *rifiuto biodegradabile* 'biodegradable waste' (+48) In particular, it appears mostly in MUD documents (*Modello Unico di Dichiarazione Ambientale* 'unified model for environmental declaration'), inside the EWC code[6] 20.01.08 *rifiuti biodegradabili di cucine e mense* 'biodegradable kitchen and canteen waste'.

While no term expressing the concept <*rifiuto indifferenziato*> 'mixed waste' is particularly associated with administrative acts, in informative texts it is mostly referred to as *indifferenziato* 'undifferentiated waste' (+179) and *secco residuo* 'dry residual waste' (+49). In technical-operational texts, the preferred term is *rifiuti urbani non differenziati* 'general mixed waste' (+23), which corresponds to the EWC code 20.03.01.

[6] European Waste Catalogue

Subcorpus	VdB	fu	hu	ha	TTR
admin	41.9%	71.1%	23.1%	5.7%	0.79
tech	36.0%	71.6%	22.6%	5.8%	0.80
info	37.6%	67.8%	23.0%	9.3%	0.70
baseline (itTenTen)	60.3%	73.9%	22.4%	3.8%	0.74

Table 2: Type-token ratio (TTR) and percentage of words from the basic vocabulary (Vdb), further divided by repertoire of use. I.e., fundamental (fu); high-usage (hi); and high-availability (ha).

The concept <*rifiuti urbani*> 'municipal waste' is expressed in administrative acts mainly through the phrase *rifiuti solidi urbani* 'municipal solid waste' (+58) and the acronyms *RU* (+34) and *RSU* (+24). No term expressing this concept has a positive relationship with informative texts while technical-operational texts are strongly associated with the term *rifiuti urbani* 'municipal waste' (+667).

The concept <*raccolta porta a porta*> 'door-to-door waste collection' is mostly conveyed in administrative acts through the terms *raccolta domiciliare* lit. 'domestic collection' and *servizio di raccolta domiciliare* lit. 'domestic collection service'. Conversely, in informative texts, the preferred variants are the terms *porta a porta* 'door-to-door' (+177) and its acronym *PAP* (+142).

5. Conclusion

Socio-pragmatic uses of institutional Italian comprise special and media institutional languages. The former is used to legislate and administrate and the latter to communicate with the general public through various media: newspapers, websites and advertising material. For these uses, institutional Italian has different lexica and employs different terms, with various degrees of specialization, to refer to similar concepts.

In order to define the lexical profile of institutional Italian, we collected 82 different terms expressing 6 concepts and examined their distribution across the three subcorpora of the ItaIst-DdAC_GRU corpus, namely administrative acts, informative texts and technical-operational texts.

Results show that administrative acts employ high-specialization terms compliant with the law, often in the form of acronyms. Conversely, informative texts contain more low-specialization terms and make extensive use of single-word terms and acronyms to remain self-contained. The terminology of technical-operational texts is largely composed of standardized and formulaic phrases.

Furthermore, results also suggest that standard metrics of lexicon complexity that do not consider terminology may lead to erroneous conclusions when applied to specialized corpora and should therefore be carefully interpreted and preferably complemented with the analysis of terminological variation.

In the future, we aim to develop an index of terminological specialization and a method to accurately measure the lexical and terminological complexity of specialized corpora.

6. Acknowledgements

This work was conducted within the PRIN 2020 Project *VerbACxSS: su verbi analitici, complessità, verbi sintetici, e semplificazione. Per l'accessibilità*, funded by Ministero dell'Università e della Ricerca (PRIN 2020_2020BJKB9M).

7. Bibliographical References

Gaetano Berruto. 1987. *Sociolinguistica dell'italiano contemporaneo*. Carocci. 2ª ed. 2012.

Maria Teresa Cabré. 1999. *Terminology: Theory, methods, and applications*, volume 1. John Benjamins Publishing.

Michele Cortelazzo. 2021. *Il linguaggio amministrativo. Principi e pratiche di modernizzazione*. Carocci.

Tullio De Mauro and I Chiari. 2016. Il nuovo vocabolario di base della lingua italiana. *Internazionale*.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

Annibale Elia, Alessandro Maisto, Lorenza Melillo, and Serena Pelosi. 2021. The lexical complexity and basic vocabulary of the italian language. In *Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities: 14th International Conference, NooJ 2020, Zagreb, Croatia, June 5–7, 2020, Revised Selected Papers 14*, pages 14–23. Springer.

Angela Ferrari and Filippo Pecorari. 2022. *Le buone pratiche redazionali nei testi istituzionali svizzeri in lingua italiana*. Franco Cesati.

Alfredo Fioritto. 1997. *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*. Il Mulino.

Riccardo Gualdo and Stefano Telve. 2011. *Linguaggi specialistici dell'italiano*. Carocci.

M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, and V. Suchomel. 2013. The tenten corpus family. In *7th International Corpus Linguistics Conference CL*, pages 125–127.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*, volume 6.

Sergio Lubello. 2018. L'antilingua gode di buona salute: nuove forme, vecchi vizi. In *Comunicare cittadinanza nell'era digitale Saggi sul linguaggio burocratico 2.0*, pages 31–43. FrancoAngeli.

Pietro Lucisano and Maria Emanuela Piemontese. 1988. Gulpease: una formula per la predizione della leggibilita di testi in lingua italiana. *Scuola e città*, pages 110–124.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 849–866.

Lucia Passaro and Alessandro Lenci. 2016. Extracting terms with extra. In *Computerised and corpus-based approaches to phraseology: Monolingual and multilingual perspectives*, pages 188–196. Tradulex.

Maria Emanuela Piemontese. 1996. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Tecnodid.

Maria Emanuela Piemontese. 1999. La comunicazione pubblica e istituzionale. il punto di vista linguistico. In Stefano Gensini, editor, *Manuale della comunicazione*. Carocci.

Tommaso Raso. 2005. *La scrittura burocratica. La lingua e l'organizzazione del testo*. Carocci.

Luca Serianni. 2007. *Italiani scritti*. Il Mulino.

Alberto A. Sobrero. 1993. *Introduzione all'italiano contemporaneo. La variazione e gli usi*. Laterza.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *CLiC-it/EVALITA*, pages 4333–4338.

Daniela Vellutino. 2018. *L'italiano istituzionale per la comunicazione pubblica*. Il Mulino.

Daniela Vellutino and Nicola Cirillo. 2024. Corpus «itaist»: Note per lo sviluppo di una risorsa linguistica per lo studio dell'italiano istituzionale per il diritto di accesso civico. [Manuscript submitted for publication].

Daniela Vellutino, Federica Marano, and Annibale Elia. 2012. L'italiano istituzionale e le sue varietà d'uso pubblico. aspetti lessicali nei tipi di testo d'informazione e comunicazione delle pubbliche amministrazioni. In *Atti XI Congresso SILFI*.

8. Language Resource References

Passaro, Lucia C. and Lenci, Alessandro. 2019. *PaWaC - Public Administration Web as Corpus Corpus*. distributed via European Commission, Directorate-General for Communications Networks, Content and Technology. PID http://data.europa.eu/88u/dataset/elrc_1282.

Tonelli, Sara and Palmero Aprosio, Alessio and Saltori, Francesca. 2019. *SIMPITIKI corpus for simplification in Italian*. distributed via Zenodo. PID https://doi.org/10.5281/zenodo.2535632.

A. Terms selected for the study

Centro di raccolta CCR; CdR; centro comunale di raccolta; centro comunale di raccolta rifiuti; centro di raccolta; centro di raccolta comunale; centro di raccolta dei rifiuti urbani; centro di raccolta intercomunale; centro di raccolta rifiuti; centro di raccolta rifiuti solidi urbani; centro di raccolta rifiuti urbani; centro di raccolta temporaneo; CRC; eco isola; ecocentro; ecocentro comunale; ecopiazzola; isola ecologica; isola ecologica comunale; isola ecologica itinerante.

Rifiuto organico FORSU; frazione biodegradabile; frazione organica; frazione organica di rifiuti; frazione organica umida; frazione umida; organico; rifiuto biodegradabile; rifiuto organico; rifiuto umido; umido.

Rifiuto indifferenziato frazione indifferenziata; frazione indifferenziato residuale; frazione non riciclabile; frazione residua; frazione rifiuti indifferenziati; frazione secca indifferenziata; frazione secca non differenziata; frazione secca non riciclabile; frazione secca residua; frazione secca residua indifferenziata; indifferenziato; materiale non riciclabile; residuo indifferenziato; residuo secco; rifiuti domestici indifferenziati; rifiuti urbani indifferenziati; rifiuti urbani non differenziati; rifiuto indifferenziato; rifiuto indifferenziato residuale; rifiuto residuo; rifiuto secco indifferenziato; rifiuto secco non riciclabile; rifiuto secco residuo; RSU indifferenziati; secco indifferenziato; secco non riciclabile; secco residuo; secco residuo indifferenziato.

Rifiuti urbani RSU; RU; rifiuti solidi urbani; rifiuti urbani.

Raccolta differenziata differenziata; raccolta differenziata; raccolta differenziata dei rifiuti; raccolta differenziata dei RSU.

Raccolta porta a porta PAP; porta a porta; raccolta differenziata domiciliare; raccolta differenziata porta a porta; raccolta domiciliare; raccolta porta a porta; raccolta rifiuti porta a porta; servizio di raccolta domiciliare; servizio porta a porta; sistema di raccolta differenziata porta a porta; sistema di raccolta domiciliare; sistema porta a porta.

Concept	**Term structure**		
	sw	mw	acronym
Centro di raccolta waste *recycling centre*	2	15	3
Rifiuto organico *organic waste*	2	8	1
Rifiuto indifferenziato *mixed waste*	1	28	0
Rifiuti urbani *municipal waste*	0	2	2
Raccolta differenziata *separate collection*	1	4	1
Raccolta porta a porta *door-to-door waste collection*	0	11	1
TOT	6	68	8

Table 3: Terms selected for the study, divided into single-word terms (sw), multi-word terms (mw) and acronyms.