Prethodno priopćenje

UDK 81’243.3(450)
Prethodno priopćenje
Prihvaćeno za tisak: 20. rujna 2019.
https://doi.org/10.22210/suvlin.2019.088.07

Marta Ghilardi
Libera Università di Bolzano
maghila84@gmail.com

Eliciting comparable spoken data in minor languages: first observations from the corpus Kontatti

In this contribution, we will deal with the issue of building a spoken corpus of conversational data that can be easily compared across languages. We will present linguistic codes embedded in Trentino and South Tyrol, where multilingualism (*de jure*) is the rule. In this area of northern Italy, more than two languages and cultures coexist and are in contact with one another. The corpus includes the major languages Italian and German and minor languages and dialects belonging both to the Romance language group, such as Ladin and Trentino dialect varieties, and from the Germanic language group, such as Cimbrian and South Tyrolean dialects. We will discuss the methodology used to elicit spontaneous spoken data in minor languages and dialects, focusing on the Map Task (Anderson et al. 1991), which has been shown to be an efficient technique for eliciting semi–spontaneous dialogues and providing a representative sample of pragmatic, textual and syntactic contexts that are at least partially expected, and thus comparable (Cerrato 2007: 9). This technique allows the speakers to focus on extra–linguistic context and on a problem–solving task, reducing both the observer’s paradox and the speakers’ monitoring of their linguistic production.

1. Corpus Linguistics and minor languages: an introduction

The origin of corpus linguistics can be traced back to the publication, in 1964, of the Brown Corpus, which contains a million words of written American English. In the 1960s, corpus linguistics was based on written language, and was therefore disconnected from social variables and sociolinguistics. Many things have changed since then, and corpus linguistics today is applied to many different research contexts and also allows interdisciplinary studies (Baker 2011: 18), which are very important for explaining corpus data. Corpora are tools of language description, which can be best analysed through the lens of different disciplines, including sociology, sociolinguistics, history etc. There is still debate about whether corpus linguistics should be defined as a theory or a methodology (McEnery et al. 2006: 7–8). We argue that corpus linguistics is a methodology which can be applied to the analysis of every level of language, and connected to other areas of research, such as language acquisition, teaching, and other academic fields.

Nowadays, collections of (semi–)spontaneous spoken language data, also defined as corpora of spoken and conversational data, have become pivotal both in monolingual and multilingual linguistic analyses,[1] in that they can be structured to represent both a (variety of a) language, and the complexity of related repertoires, providing information on the sociolinguistic community using that particular language.

Although for some major languages, such as English,[2] there is already a conspicuous amount of data, and thus, a certain degree of representativeness, corpus linguistics has three challenges for the future: i) building corpora of endangered and minor languages in order to document them, and make the preservation of their cultural values possible, ii) creating bi— and multilingual corpora; iii) trying to collect and relate sociolinguistic information to linguistic features. In this contribution, we will deal in particular with the first two challenges. After a sociolinguistic description of the area in which we conducted the research, Trentino and South Tyrol (§2), we will describe an example of a multilingual corpus (§3), and present some data (§4) to show the potential of corpus analysis at different levels of language, and its use in collaboration with other academic fields, such as teaching.

2. Trentino and South Tyrol: a multilingual area

The area of research is the Italian region Trentino–Alto Adige/Südtirol, which includes two provinces: Trentino (officially, the province of Trento) and South Tyrol (officially, the province of Bozen–South Tyrol). It is a multilingual region and represents a complex geographical area from a sociolinguistic point of view (Mioni 2000). For this reason, we decided to collect data in this context, where the building of a multilingual corpus is a real challenge for both corpus linguistics and sociolinguistics.

In the province of Trento, Italian is spoken, and, in addition, three minor languages — Ladin, a Romance language, and Cimbrian and Mocheno, Germanic languages.[3] The latest census[4] reports that the Ladin population amounts to 3.5% of the total population of the province of Trento; speakers are located in particular

[1] See the contributions in Schmidt and Wörner (eds.) (2012).

[2] For an overview, see https://www.english–corpora.org/

[3] http://www.minoranzelinguistiche.provincia.tn.it/minoranzeTrentino/

[4] http://www.statweb.provincia.tn.it/pubblicazioniHTML/Annuari%20e%20altre%20pubblicazioni%20di%20carattere%20generale/Annuari%20statistici/Annuario%20statistico%202011/AnnStat2011.pdf and http://www.statistica.provincia.tn.it/binary/pat_statistica_new/popolazione/RilevazionePopolazioniLadinaMochenaCimbra.1394031752.pdf

in the Fassa Valley. The highest number of Cimbrian speakers is settled in Luserna and amount to 0.2% of the total population of the province of Trento. The Mocheno–speaking population, concentrated in Palù del Fersina and Fierozzo, amounts to 0.3%.[5]

The language policy of South Tyrol, in force since 1972, officially recognizes Italian, German and Ladin. However, as described in Dal Negro and Ciccolone (2018), in South Tyrol the multilingual situation is more complex than it seems. Alongside a *de jure* multilingualism, which encourages the use of these three languages in public settings, there is a *de facto* multilingualism (Dal Negro 2017). If we focus on the latest census report for the population of the province of Bozen–South Tyrol, we see that the linguistic groups are divided almost homogeneously in different areas: the German majority is located in particular in the mountains; Italian speakers live in the city of Bozen–Bolzano; and Ladin speakers are concentrated in the Ladin valleys Gardena and Badia. While in the cities there is more opportunity to use the different languages, in the mountains this tends to be limited to school and to communication with tourists. We can also observe that 69.41% of the population of South Tyrol declares itself as belonging to the German linguistic community, 26.06% as belonging to the Italian linguistic community, and the remaining 4.53% to the Ladin linguistic community. Parameters like the distribution of the members of the different linguistic groups in the territory and the history of the regions — which shape language attitudes and language policy — influence the real degree of bilingualism of the inhabitants of this area.

It is important to emphasise that alongside the minor and major languages spoken in Trentino and South Tyrol, dialects[6] also play a role in the repertoire of the communities. In particular, the Bassa Atesina community in South Tyrol displays the use of two dialects: Trentino (an Italo–Romance dialect) and Tyrolean (a German dialect), frequently mixed and used, as we will see in the sections below, as a 'we–code', an intra–group code with identitarian value (Dal Negro 2018: 74–76).

3. The corpus

The corpus was created as part of the Kontatto project.[7] More than 80 speakers, aged between 13 and 81 years, were interviewed, in the geographical area of Bassa

[5] See national law 482/1999 for the protection of the linguistic minorities: http://www.minoranzelinguistiche.provincia.tn.it/normativa/Normativa_nazionale/pagina5.html this law guarantees the safeguard of the minor languages, in particular through the teaching of them in the schools, permanent education programs, and university courses; the administrative apparatus is also allowed to use the minor languages, alongside Italian. The law for linguistic minorities in the province of Trento can be accessed at: http://www.minoranzelinguistiche.provincia.tn.it/binary/pat_minoranze_2011/NormativaPAT/Legge_provinciale_7ago2006_numero5_al_31dic2016.1485344775.pdf

[6] With the term 'dialect' we do not intend varieties of the major language, but autonomous languages.

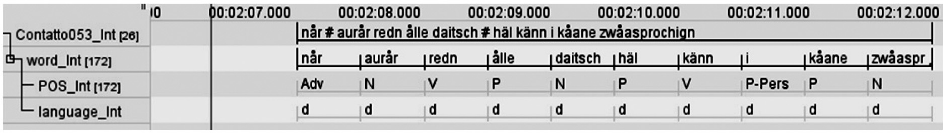
[7] The project was funded by the Province of Bozen–Bolzano and coordinated by Prof. Silvia Dal Negro. More detail can be found at: http://kontatti.projects.unibz.it/

Atesina in South Tyrol. Around 18 hours of spoken interactions were recorded, using different methodologies, such as the Map Task, interviews, and spontaneous speech among family members and friends. All speakers filled out a sociolinguistic questionnaire, in either Italian or German.

During a subsequent project, called Kontatti, the corpus was enlarged, investigating a wider geographical area that extends from South Tyrol to Trentino, through Ladin valleys and the Cimbrian territory. In particular, 50 speakers (between 16 and 97 years old) participated in these interviews for an overall 6 hours and 48 minutes of recordings. The methodology used to create the two parts of the corpus was fundamentally the same. Before accomplishing the map task, the speakers had to fill out a consent form, and a sociolinguistic questionnaire. As regards the latter, they could choose between the standard versions of the languages spoken in the different parts of the investigated territory (German, Italian, Fassa Ladin, Gardena Ladin, Cimbrian).

From a methodological point of view, the importance of the compilation of a questionnaire lay in the fact that, on the one hand, the speakers had time to make themselves "comfortable", and become accustomed to the recorder, and on the other, researchers could also record spontaneous speech, which could be analysed after transcription. It was felt that if all interviewers were part of the linguistic community in which they carried out the interviews it would hopefully reduce the observer's paradox phenomenon, and made it possible for the informants to use all the linguistic codes of the community, in particular dialects and minor languages, based exclusively on the linguistic community's rules, without any limitation imposed by external factors, such as a potential stigma towards some of the varieties in the repertoire.[8] All the interviews were transcribed using Elan.[9]

The data in the corpus were treated differently.



00 00:02:07.000 00:02:08.000 00:02:09.000 00:02:10.000 00:02:11.000 00:02:12.000

Contatto053_Int [26] når # aurår redn ålle daitsch # häl känn i kåane zwåasprochign

word_Int [172]	når	aurår	redn	ålle	daitsch	häl	känn	i	kåane	zwåaspr
POS_Int [172]	Adv	N	V	P	N	P	V	P-Pers	P	N
language_Int	d	d	d	d	d	d	d	d	d	d

Figure 1. Data treatment of Kontatto

In Figure 1, which shows an excerpt from the Kontatto corpus, we can observe an orthographic tier, followed by a *word* tier where the utterance is automatically divided into words; the *POS* tier is dedicated to the part of speech tagging, aligned with the relative language (*language* tier). Both the POS— and the Language–tags[10] include the categories proper noun and interjections. All the words that could not be ascribed to a specific language were tagged as such; in addition to onomastic

[8] About the prejudices towards the Italian dialects, see Ruffino (2006), Marcato (2007).

[9] https://tla.mpi.nl/tools/tla–tools/elan/

[10] Transcription, POS and Language tags were processed manually.

references, examples of this include proper nouns of place, such as *Bar Sport*, and discourse markers such as *mh, eh,* etc. Exception was made for some toponomastic references such as Alto Adige, the name in Italian for the province of Bolzano, or Südtirol, the German equivalent: in such cases the language choice of the speakers can cast some light on the language attitudes.[11]

By comparison, in Figure 2, we can observe the data treatment in the Kontatti corpus, organised at the level of utterances, rather than parts of speech as in the Kontatto corpus.

We have an orthographic tier, the subject expression tagging, and the relative possible language/s: German, Italian, Gardena Ladin, Fassa Ladin, Cimbrian, Trentino, Tyrolean dialect, and mix when the utterances contain code mixing.[12]

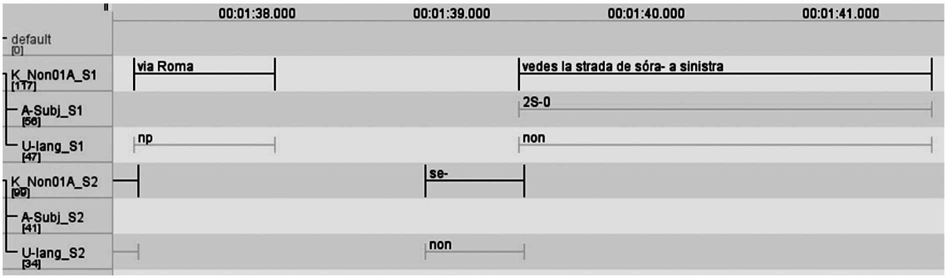


Figure 2. Data treatment of Kontatti

An additional tier — the *V2* tier — was created for Ladin varieties in Kontatti (Figure 3), in order to investigate the syntactic contexts in which inversion between verb and subject occurs. As we will see in §4 (ex 12, 13, 14), Ladin is a Romance language but tends to follow the V2 rule typical of German languages. In particular, the verb in the clause must be in the second position in the syntactic context of direct questions, along with the inversion of the subject, for both Fassa and Gardena Ladin varieties and, for Gardena Ladin, in the presence of a topicalized element in first position. We distinguished among three parameters: XVS (inversion realized where expected); VS (inversion realized, but not expected); XSV (inversion not realized where expected):

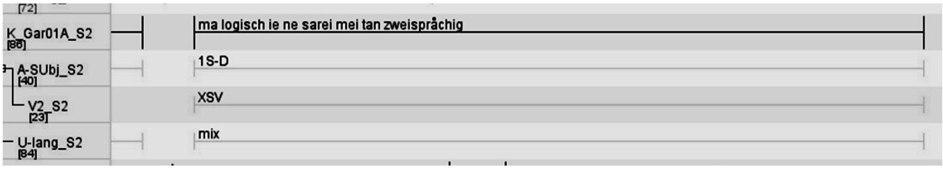


Figure 3. Data treatment in Kontatti for Fassa and Gardena Ladin

[11] Currently, we are working on the lemmatization process.

[12] In this contribution, under 'code mixing', we intend each alternation between the languages.

Finally, through the sociolinguistic questionnaires, we created a metadata section in Microsoft Excel. In particular, the following variables were recorded: language choice of the questionnaire, age, sex, place of residence, education, informants' jobs, the L1 of their parents, the first/second/third language acquired, the relationship between languages and domains of communication, questions to elicit language attitudes, such as: "Whom would you not speak Italian (dialect) with?" and "Do you consider yourself to be bilingual?".

The data in Konatti show an equal distribution in the choice of the language questionnaire by the 12 speakers of the Bassa Atesina: 6 speakers opted for Italian and 6 for German. The two official languages, Italian and German, were stated by the participants to be the first languages that they had acquired, and in general as the languages they can speak better; only two speakers stated they can speak Trentino dialect better than Italian. Nevertheless, German and Italian are limited to high domains of communication, such as in communication with teachers and professors at university and school, at work, with tourists, and, in general, with people who cannot understand dialects and who are perceived as elements that are external to the community. Tyrolean and Trentino dialects are the linguistic codes that informants declare using with parents (8 informants), grandparents (8 speakers), siblings (8 informants), neighbours (8 informants); in 6 cases dialects are also used in shops. Two speakers declared that they would not use Italian with their grandmother, because of the resentment towards Italian for political and historical reasons. One speaker also declared that he would not use dialect with "Italians who believe themselves to be better than people who speak dialect".

From these answers we can infer two elements in particular: i) German and Italian are perceived as languages with high prestige, since they are both used in high domains of communication, but some informants also relate Italian to negative attitudes, due to the fact that South Tyrol was forced to become an Italian territory after the First World War; ii) dialects are linguistic codes with identitarian values, often stigmatized by people who are not members of the community.

In terms of the Ladin community (we collected and analyzed data from the Gardena Valley in South Tyrol and the Fassa Valley in Trentino), the 6 speakers from Fassa chose the questionnaires in Fassa Ladin variety, although the first language they declared to have acquired was Italian in 3 cases, and Ladin in the remaining 3. Ladin was reported by all speakers to be the language used with family members, grandparents and neighbours; Italian was reported as being used in addition to Ladin in friendship domains, at school/university, and at work. It is important to emphasise that all speakers declared that they would never use Italian with people who know Ladin, family members and with people from Fassa.[13]

As for the speakers from Gardena, 5 chose Gardena Ladin and 1 chose German. The informants from Gardena stated that the language they acquired as a first language and speak better was Ladin (only one chose German), followed by German,

[13] See Dell'Aquila & Iannaccaro (2006).

and then Italian. These informants also stated that they would never use Italian with: people from Gardena, family members and friends, an informant added: "also if I could speak Italian with them". The latter answer testifies that apparently proficiency in the official language does not play a role in language choice.

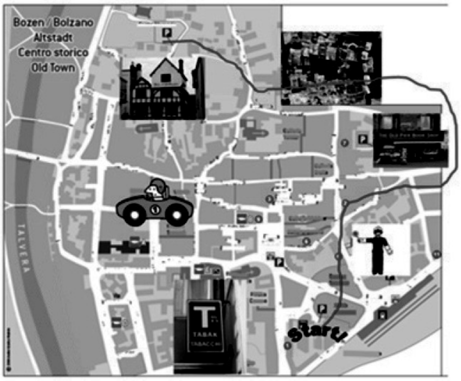
The Cimbrian language also seems to be perceived as a language with an identitarian value, with all 8 informants describing it as the language used with family. Nevertheless, 5 chose the questionnaire in Italian, compared to 3 in Cimbrian. 4 indicated Cimbrian as the first language they had learnt, 5 declared it to be the language they speak best, but only 3 chose it as the language of the questionnaire. All speakers stated that they would never use Italian with family and people from Luserna.

This kind of sociolinguistic information, along with the linguistic data in section §4, enables us to draw more accurate conclusions about the use and the role of the languages spoken in the territories of Trentino and South Tyrol, casting light in particular on minor languages such as Cimbrian, Ladin, and also dialects.

4. Data

In this section, we will present some uses of the corpus, applied to different levels of analysis. The multilingual dimension of our corpus is related to the fact that we interviewed plurilingual people, who could choose which language they preferred to use. They did not translate texts or lists of words, as in linguistic atlases that analyze the same area, such as Vivaldi or Ald–I, Ald–II.

In order to both obtain comparable data and preserve the spontaneity of the natural speech, we chose to apply the 'Map Task' technique, which was developed by the HCRC Map Task team at Edinburgh (Cerrato 2007). This technique is an important way of eliciting dialogic speech. Only two participants are involved, each of whom has a map, but can see only their own. During the first turn, one of the two, the instruction giver, has a map with a path on it, and s/he has to guide the other participant, the instruction follower, from point A to point B. Then the roles switch.



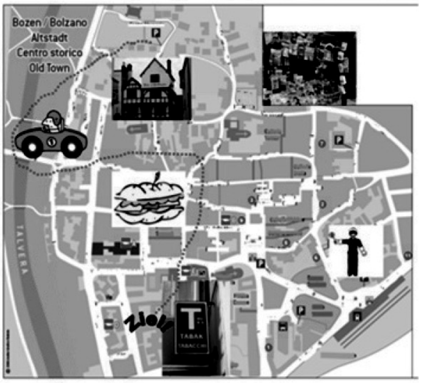


Figure 4. Map Task A Figure 5. Map Task B

The peculiarity of this technique lies in the fact that the maps are a little bit different from one another, which triggers conversation, for example clarification requests aimed at avoiding misunderstandings (Filipi 2014: 366–367). This feature is especially significant in the case of recordings of bi— or multilingual speakers, where the search for agreement can trigger the use of a different linguistic code, and even more significant when the corpus consists of minor languages or dialects, which tend to be stigmatized and used only in informal contexts:

(1) *K–Lag01_S1: e ànca àla mia destra probabilmente perché ghè — ghèra en segnale lì*
‘Also on my right probably because there was a signal there’
K_Lag02_S2: non ho capito niente sai no
‘I haven’t understood anything, you know’

In the example above (1) the instruction giver (S1) speaks Trentino dialect, while the instruction follower (S2), apperently in order to express disappointment, answers in Italian.

In the excerpt below (2), by comparison, the instruction follower (S2) uses Trentino dialect, while the instruction giver (S1), in order to answer, starts the turn of conversation in Italian: *questo* ‘this’, before switching immediately after to Trentino dialect:

(2) *K_Lag02_S2: té té té gài fàto su en gra gran caos*
‘You have done a mess’
K–Lag01_S1: ***questo*** *lè el mé percorso*
‘That is my path’

The spoken data elicited through a map task are also comparable because the speakers have to give instructions using the references on the maps. This allows us to focus on a certain number of target words. The following examples (in which the main language of communication is Tyrolean dialect) from the corpus provide valuable insights at the level of lexicon:

(3) *K012_S1: Sigsch T fån dår* ***tabaktrafik****?*
‘Do you see the T of tobacconist’s?’
(4) *K012_S1: Unt dårnoch foorsch äh ban schilt* ***tabak*** *net*
‘And then go past the drugstore sign’
(5) *K012_S1: Ja,* ***tabacchi****, genau*
‘Yes, drugstore, exactly’
(6) *K037_S1: Häm foorsch når # or bis wo ‘s foto fån* ***tabacchino*** *unheb*
‘There you go then # down where the drugstore sign starts’
(7) *K037_S1: In* ***tabacchin*** *do schun hån i gsegn*
‘The drugstore yes, I have seen it’

(8) *K012_S2:* *Rächts # fån* ***tabakilä***
‘On the right # at the drugstore’

In these examples the main language of interaction is Tyrolean and all the informants come from Bassa Atesina. In example (3), the word *tabaktrafik* comes from the Austrian German variety; the excerpt in (4) presents the German equivalent, *Tabak*; the same speaker then uses, in (5), the Italian insertions *tabacchi,* and, in (6), *tabacchino*. This is followed by an insertion from Trentino dialect, *tabacchin*, in (7), while example (8) concludes with the Tyrolean *tabakilä,* where the bound diminutive morpheme *–ilä* seems to be a structural calque of the bound morpheme *–ino* from Italian and *–in* from Trentino dialect, found in *tabacchino* and *tabacchin* respectively, although in these cases, the two bound morphemes do not have a diminutive but an agentive function.

The Bassa Atesina community has at its disposal all the linguistic codes used in the area; this is shown on the one hand by the use of all of them during the map task, and on the other by the language attitudes of the speakers, which were recorded both through the interviews and the questionnaires:

(9) *K019_S2:* *so a mischwarietet isch ainzigartig?*
‘So a mixed variety is fantastic?’
K019_S1: *Ja wail do ba ins # es gib a af nirgnds åndårsch af dår gånzn wält lait dass so redn wia mir*
‘Yes, because here among us there is nobody in the world that speaks like us’
K019_S2: *also kånnsch sogn mir kännän schtolz drauf sain*
‘So, you can say we can be proud of it’

In the examples below (table 1), it is instead clear how a multilingual corpus can also cast some light on the switch as a turn taking device,[14] in other words, the alternation between two languages signalizes a change of turn between the speakers. The importance of the map task technique in this case is related to the fact that it allows us to analyze the same conversational context.

In table 1, we can observe the start of the turn of the conversation. We analyzed the starts of conversational turns, starting with the word ‘so’ with a discourse marker function: *also* (German), *allora* (Italian); *alóra* (Trentino dialect); *bën* (Gardenese Ladin); *dapò* (Fassa Ladin):

[14] See Sacks, Schegloff and Jefferson (1974) for one of the first models of turn taking organization for conversation.

Main language of interaction	Switch	no.	Example
German	no	17	***also*** *iå ebm mir sain afn parkplåtz* ‘so yes, I am at the carpark as well'
German	Italian/ Trentino	5	***alóra*** *du bisch schtartäsch fo dän hotäl då* ‘so, you are you start from that hotel there’
Trentino	no	3	***alóra*** *da lì té vai én Zó dé dói* ‘so, from there you go on two more'
Trentino	Italian	3	***allora*** *té vègni av— té vègni avanti dé dó quadri* ‘so, you come ahead two boxes'
Italian	no	4	***allora*** *# partiamo dalla zona numero uno* ‘so # let’s start from zone number one'
Gardena Ladin	no	5	***bën*** *po l A chël ie ehm a man ciancia dla plata gonz su insom* ‘good, then A that is ehm on the left of the sheet, up basically'
Gardena Ladin	German	2	***also*** *## tu es pu la Piaz* ‘so ## you are at the square'
Fassa Ladin	Italian	5	***allora*** *parte dal prum # riquadro* ‘so, it starts from the first # box'

Table 1. Code–switching as a turn taking device

We can observe that, when the main language of interaction is German, in 17 cases we do not have any switch, but in 5 cases the switch is to Italian/Trentino; when the main language of interaction is Trentino dialect, we observe in 3 cases a switch to Italian and in 3 cases there is no switch; in the interactions in which Italian is the main language, there are no switches; in Gardena Ladin, we can observe switches to German in two cases, while in 7, there is no switch strategy; when Fassa Ladin is the main language of communication the switch tends to be to Italian. The standard Fassa Ladin words for ‘so’ are *enlouta, dapò, embendapò* and *emben*, but in the corpus we can observe only *dapò,* and this seems to have maintained the meaning of the temporal adverb ‘then’. Elsewhere, the Italian form *allora* has taken over the function as a discourse marker (Fiorentini 2014: 98–101).

Overall, Italian seems to be the most switched to language. Gardena Ladin speakers are more inclined to switch to German, whereas the Fassa Ladin speakers prefer to switch to Italian; Ladin is never switched to. Cimbrian does not appear in table 1 because it uses always the Trentino word ‘*alóra*’, but it has been officially recognized as a loan word.

A preliminary analysis of Kontatti corpus shows that we tagged 1004 utterances at language level in the recordings from the Gardena valley. Of these, 822 utterances are in Gardena Ladin, 42 in German, 9 in Italian, while in 131 we find code mixing. In particular, there are 18 Ladin–Italian utterances, 3 Ladin–Italian–

German utterances, and 110 Ladin–German utterances. One such example is as follows:

(10) *K_Gar01B_S2: na* ***kreizung*** *chësc ie mé na streda che va ju.*
‘a **crossroad**, this is just a street that goes down’

In the example above (10) the main language is Ladin, while the insertion *kreizung* (crossroad) is in Tyrolen dialect. *Crossroad* is also a target word in the map task; it appears 8 times in the form *ncrëusc,* which like the Italian *incrocio*, is masculine, and 12 times in the form *kreizung*, which is feminine. The standard Gardena Ladin word for ‘crossroad’, however, is *ncrujeda,* which is feminine. The form *ncrëusc* seems to be a morphologic calque of Italian. Even though the speakers do not hold positive attitudes towards Italian, Italian is not a favoured language in switches and only 9 utterances in Italian are recorded, the contact between Italian and Gardena Ladin also seems to have had consequences at a deeper level than the lexicon, a level that cannot be easily controlled by the speakers.

In the 625 utterances in the Fassa Ladin analyzed, we found no utterances in German (this was as expected, as German is not an official language in the province of Trento), 52 in Italian, 422 in Ladin, and 151 mixed with Italian. Returning to the target word *crossroad,* we find the Italian *incrocio* 6 times, and only once the Fassa Ladin *crousc de via*. In this case also, Italian appears to be strongly present in the Ladin variety of our speakers.

The above examples of analysis relate to code mixing applied to lexical and conversational analysis, allowing us to draw a more accurate picture of the relationship among the languages in the repertoire. The following example of investigation, by contrast, relates to the relationship between corpus linguistics and teaching.

We are currently analysing the expression of the subject in Gardena Ladin in order to compare the spoken language with the grammar books for Gardena Ladin used at school (Ghilardi and Videsott, in print), and support the teaching of a minor language. In this area of analysis, the map task was again useful, as the dialogic form of the task, based around instruction giving, implemented the use of subject pronouns, in order to explain who does what. This can be seen in example (11) below, where the instruction follower Gar02B_S2 asks for confirm of his/her next move, using the stressed form of subject pronoun of the first person singular (1ST) ie (I), and the instruction giver Gar02B_S2 explains what he/she has to do, using the stressed form of the second person singular (2ST) tu (you):

(11) *Gar01B_S2: praktisch* ***ie*** *# ie ved via # a man drëta?*
‘Basically I(1ST) # I(1ST) go away # to the right’
Gar01B_S1: ***tu*** *ves via chin te chël ncrëusc ulach — # tl prim ncrëusc*
‘you(2ST) go till that croassroad where — # the first croassroad’

So far, we have analysed the different forms of the first person singular, in order to observe the relation between syntactic context and the two morphological forms:

	Stressed	Unstressed (procl/encl)
1st person	ie	Ø/–i

Table 2. Subject pronouns: 1st person singular, Gardenese Ladin

The grammar books of Gardena Ladin state that the enclitic subject pronoun *–i* (1SCl) occurs only in inversion contexts XVS (in direct questions and V2 context; in other words, the subject follows the verb when there is a topicalized element in first position); while the stressed form (1ST) and the null subject (1SØ) are free variants in SVO contexts, but the stressed form is the favourite choice when the speakers have pragmatic intents.

We analyzed the utterances on the basis of three possible contexts: XVS (inversion realized where expected); VS (inversion realized but not expected); XSV (inversion not realized where expected):

(12) XVS

Gar01B_S1:	*tlo*	*scrij–**i***	*mi*	*inuem*
	ʻhere	write–I(1SCl)	my	name’

In this example (12) the enclitic form of the subject pronoun of the first person singular –i (1SCl) agrees with the descriptions in the grammar books. The inversion between subject and verb is required, because of the presence of the adverb *tlo* (here) in first position.

(13) VS

Gar01B_S1:	*ved–**i***	*via*	*n*	*doi*	*chedri*	*eh!*
	‘go–I(1SCl)	away	a	two	squares	eh!’

In example (13) above, we observe the inversion of the subject, hence the use of the first singular person enclitic form *–**i*** (1SCL), in an SVO context. The grammar books, however, suggest the use of the stressed form ***ie*** (1ST), or the null subject ***Ø*** (1S Ø) in such a context: *ie/Ø vede via n doi chedri eh!*

(14) XSV

Gar02A_S1:	*Normal*	***ie***	*ved*	*for a Trënt!*
	ʻUsually	I(1ST)	go	to Trento’

Excerpt (14) shows the stressed form of the subject pronoun of the first person singular ***ie*** (1ST). The grammar books, however, require the unstressed form –i (1SCl), because of the presence of the frequency adverb *normal* in first position: *Normal ved–i for a Trënt* .

It is also interesting to observe that the verb *ved*— (to go) lacks the inflectional suffix *–e: ved–e,* as should be the case for first person verbs in Gardena Ladin (*ie vede* > I go).

The XSV context with the first person singular, however, appears only five times, because the clitic form of the subject of the first person singular ***–i*** is extended to every syntactic context, and is not limited to the inversion context.

The clitic form of the subject of the first person singular ***–i*** grammaticalized in the verbal inflectional suffix of the first person singular is said in the grammar books to be *–e*. However, our spoken data allow us to conclude that the inflectional suffix of the first person singular *–e* is completely substituted by *–i*.

5. Conclusion

The main points discussed in the present contribution outline the creation of a corpus which attempts to answers several challenges of corpus linguistics, in particular i) the creation of a multilingual corpus, where the multilingual dimension lies not in the translation of texts or words, but in the recording of plurilingual speakers, in order to cast some light in the functioning of both individual and community repertoires; ii) the recording of minor and endangered languages, in order to also meet preservation aims.

Our data were collected in a geographical area, Trentino–Alto–Adige/Südtirol, where multilingualism is guaranteed by a language policy which recognizes Italian, German and Ladin in South Tyrol; in Trento Italian, Ladin and Cimbrian. In addition, other linguistic communities are also present in the region, such as in Bassa Atesina (South Tyrol — providence of Bolzano), where two dialects, Trentino and Tyrolean, are used alongside German and Italian.[15]

In order to collect data that can represent a multilingual community, we applied a complex methodology: a questionnaire in each language spoken by the community (excluding Tyrolean and Trentino dialects, which do not have any official written form), which the speaker could choose (and whose compilation was recorded); spontaneous speech; and the recording of the map task. All the interviews were conducted by a member of the linguistic community. As we showed in §4, the use of the map task is, in our opinion, the best strategy for preserving the spontaneity of the speech and collecting comparable data. Finally, we enriched the linguistic data with the responses in the sociolinguistic questionnaires, relating them to the attitudes of the speakers, to draw a more accurate picture of the linguistic repertoire of the communities.

The data presented in section §4 were selected to cast some light on the different analyses that can be carried out using a multilingual corpus, at different levels of language, and that can also be applied to different branches of research, such as language learning. In particular, we presented a lexical analysis of target words in the map task, a conversational analysis focused on code switching as a turn–taking device, and an analysis of the degree of pervasiveness of the different languages in

[15] Mocheno is also a minor and endangered language in Trento province, but, so far, we have not collected any data.

the (semi)spontaneous speech of the Gardena and Fassa communities. Finally, we presented the first step of a study we are still conducting to support metalinguistic reflection on Gardena Ladin, through the investigation of the expression of subject pronouns.

All these data can allow us several observations. First, we can state that the multilingualism in the different linguistic communities we examined is not balanced.[16] Through the language tagging in the corpus, we can observe a different orientation in each community towards the languages in their repertoire. The minor languages have a strong identitarian value, and the language attitudes of the speakers help us to reconstruct a more accurate hierarchy among the languages within the different repertoires. Finally, we can also use the analysis of the data to contribute to the teaching of a minor language.

References

Anderson, Anne H. et al. (1991). The HCRC Map Task Corpus. *Language and Speech* 34: 351–366

Baker, Paul (2011). Social involvement in Corpus studies. Vander, Viana, Sonia Zyngier, and Geoff Barnbrook, eds. *Perspectives on Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing, 17–28

Cerrato, Loredana (2007). *Progetto CLIPS. Corpora e Lessici di Italiano Parlato e Scritto*. Available at http://www.clips.unina.it/it/documenti.jsp (last accessed 11/06/2019).

Dal Negro, Silvia (2018). Finding patterns in bilingual speech. *Lingue e Linguaggio* 17 (1): 71–85, https://doi.org/10.1418/90424

Dal Negro, Silvia (2017). Bilinguismo asimmetrico in Alto Adige: lo spazio sociolinguistico dell'italiano. Bombi, Raffaella, ed. *Nuovi spazi comunicativi per l'italiano nel mondo. L'esperienza di 'Valori identitari e imprenditorialità'* (Serie "Valori identitari e imprenditorialità" 4). Udine: Forum, 59–67

Dal Negro, Silvia, and Simone Ciccolone (2018). Il parlato bilingue. Calleja Bermejo, Felisa, and Peggy Katelhön, eds. *Lingua Parlata. Un confronto fra l'italiano e alcune lingue europee*. Berlin: Peter Lang, 385–407

Dell'Aquila, Vittorio, and Gabriele Iannaccaro (2006). *Survey Ladins. Usi linguistici nelle valli ladine*. Trento: Regione autonoma Trentino–Altoadige

ELAN (Version 5.2) [Computer software] (2018, April 04). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from
https://tla.mpi.nl/tools/tla–tools/elan/ (last accessed 28/06/2019)

Filipi, Anna (2014). Speakers' orientation to directional terms in a map task. *Discourse Studies*, 16 (3), 365–384, https://doi.org/10.1177/1461445613508896

Fiorentini, Ilaria (2014). Connessione e contatto. Connettivi italiani nel ladino fassano parlato. *Cuadernos de Filología italiana* 21: 85–105

[16] See Dal Negro and Ciccolone (2018).

Ghilardi, Marta, and Ruth Videsott [in print]. L'incompletezza del sistema pronominale soggetto del ladino gardenese e le sue ricadute didattiche. Dal Negro, Silvia, and Antonietta Marra, eds. *Lingue minoritarie tra localismo e globalizzazione*. Studi AitLA. Milano: Officinaventuno

Marcato, Carla (2007). *Dialetto, dialetti e italiano*. Bologna: Il Mulino

McEnery, Tony, Richard Xiao, and Yukio Tono (2006). *Corpus–Based Language Studies. An advanced resource book*. New York: Routledge

Mioni, Alberto M. (2000). La situazione sociolinguistica dell'Alto Adige/Südtyrol. Pasinato, Antonio, ed. *Heimat. Identità regionali nel processo storico*. Roma: Donzelli, 333–342

Ruffino, Giovanni (2006). *L'indialetto ha la faccia scura. Giudizi e pregiudizi linguistici dei bambini italiani*. Palermo: Sellerio

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson (1974). A Simplest Systematics for the Organization of Turn–Taking for Conversation. *Language* 50 (4): 696–735

Schmidt, Thomas, and Kai Wörner (eds.) (2012). *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam/Philadelphia: John Benjamins Publishing Company

Linguistic Atlases

[ALD–I] GOEBL H. *et al*. (eds.) (1998). Atlant linguistich dl ladin dolomitich y di dialec vejins, 1ª pert/Atlante linguistico del ladino dolomitico e dei dialetti limitrofi, 1ª parte/ Sprachatlas des Dolomitenladinischen und angrenzender Dialekte, 1. Teil/Linguistic Atlas of Dolomitic Ladinian and neighbouring dialects, 1st Part. Wiesbaden: Reichert.

[ALD–II] GOEBL H. (eds.) 2012. Atlant linguistich dl ladin dolomitich y di dialec vejins, 2ª pert/Atlante linguistico del ladino dolomitico e dei dialetti limitrofi, 2ª parte/Sprachatlas des Dolomitenladinischen und angrenzender Dialekte, 2. Teil/Linguistic Atlas of Dolomitic Ladinian and neighbouring dialects, 2nd Part, Strasbourg: Éditions de linguistique et de philologie.

[VIVALDI] Vivaio acustico delle lingue e dei dialetti d'Italia, www2.hu–berlin.de/vivaldi (last accessed 10/10/2019).

Websites

https://www.english–corpora.org/ (last accessed 28/06/2019).

http://www.minoranzelinguistiche.provincia.tn.it/normativa/Normativa_nazionale/pagina5.html (last accessed 28/06/2019).

http://www.minoranzelinguistiche.provincia.tn.it/minoranzeTrentino/ (last accessed 28/09/2019).

http://www.minoranzelinguistiche.provincia.tn.it/binary/pat_minoranze_2011/NormativaPAT/Legge_provinciale_7ago2006_numero5_al_31dic2016.1485344775.pdf (last accessed 28/09/2019).

http://kontatti.projects.unibz.it (last accessed 14/10/2019).

http://www.statistica.provincia.tn.it/binary/pat_statistica_new/popolazione/RilevazionePopolazioniLadinaMochenaCimbra.1394031752.pdf (last accessed 28/09/2019).

http://www.statweb.provincia.tn.it/pubblicazioniHTML/Annuari%20e%20altre%20pubblicazioni%20di%20carattere%20generale/Annuari%20statistici/Annuario%20statistico%202011/AnnStat2011.pdf (last accessed 28/06/2019).

Elicitazione di dati comparabili di parlato (semi)spontaneo in lingue di minoranza: prime osservazioni dal corpus Kontatti

Il presente contributo mostra l'importanza nel panorama (socio)linguistico della costruzione e produzione di corpora multilingui, riflettendo in particolare sulla registrazione di lingue minoritarie. Il corpus utilizzato per tale approfondimento teorico è stato creato dalla Libera Università di Bolzano, grazie al progetto *Kontatto: aree storiche di contatto tra Sudtirolo e Trentino*, e ampliato successivamente dal progetto *Kontatti: Discourse and structures in contact*. Il territorio interessato dall'indagine si presenta come un'area di contatto prolungato tra popolazioni, culture e identità, posta al confine tra l'area linguistico–culturale romanza e quella germanofona. I codici linguistici registrati comprendono il tedesco, l'italiano, il ladino, il cimbro e i dialetti tirolese e trentino. Il corpus è costituito da parlato (semi)spontaneo elicitato tramite la registrazione di dati durante la compilazione di un questionario sociolinguistico, conversazioni spontanee e l'utilizzo di Map task. Quest'ultima tecnica si è dimostrata particolarmente efficace per la raccolta di dati comparabili tra le diverse (varietà di) lingue, consentendoci di slegare la comparazione tra i diversi codici linguistici dalla traduzione non spontanea di uno testo. I dati prodotti da tali registrazioni ci hanno permesso, da un lato, di incrementare i dati già offerti dalla letteratura, quali, ad esempio, quelli riguardanti i repertori linguistici, riuscendo a far emergere il ruolo ricoperto dai diversi codici linguistici presenti nei territori menzionati e spiegare in modo più accurato il livello di bi–plurilinguismo delle rispettive comunità linguistiche; dall'altro di iniziare lo studio della struttura del parlato dei codici minoritari, così da applicare tale studio, onde possibile, alla didattica delle lingue minoritarie.

Parole chiave: lingue minoritarie, map task, contatto linguistico, linguistica dei corpora.

Stvaranje usporedivih govornih podataka za manjinske jezike: prvi rezultati iz korpusa Kontatti

U ovom su radu prikazani izazovi pri stvaranju usporedivoga korpusa govornog jezika. Predstavljeni su jezični kodovi iz Trentina i Južnog Tirola, gdje je višejezičnost (de jure) pravilo. Na ovom području sjeverne Italije koegzistira više od dva jezika i više od dvije kulture koje su međusobno u dodiru. Korpus uključuje glavne jezike, talijanski i njemački te manjinske jezike i dijalekte koji pripadaju romanskoj skupini, poput ladinskog i trentinskih dijalekata, kao i one iz njemačke skupine, poput cimbrijskog i južnotirolskih dijalekata. U radu se raspravlja o metodologiji koja se koristi za prikupljanje spontanih govornih podataka u manjinskim jezicima i dijalektima, usredotočujući se na zadatak karte (Anderson i sur. 1991). Zadatak karte pokazao se učinkovitom tehnikom za dobivanje djelomično spontanih dijaloga, ali i reprezentativnoga uzorka pragmatičkog, tekstnog i sintaktičkog konteksta, koji je barem djelomično očekivan, a shodno tome i usporediv (Cerrato 2007: 9). Ova tehnika omogućuje govornicima da se usredotoče na izvanjezični kontekst i na zadatak rješavanja problema, umanjujući time i promatrački paradoks i nadgledanje vlastite govorne produkcije u ispitanika.

Keywords: minor languages, multilingualism, language contact, map task, corpus linguistics, Trentino-Alto Adige/South Tyrol

Ključne riječi: manjinski jezici, višejezičnost, jezični kontakti, zadatak s kartom, korpusna lingvistika, Trentino-Alto Adige/Južni Tirol