A corpus-based analysis of referentiality in Mapudungun
Preliminary report

Lucía Golluscio
Universidad de Buenos Aires, CONICET, Argentina
https://orcid.org/0000-0001-8808-7611

Christian Lehmann
Professor emeritus at the University of Erfurt, Germany
https://orcid.org/0000-0002-9009-1916

Felipe Hasler
Universidad de Chile, Chile
https://orcid.org/0000-0003-2050-2481

Anna Pamies
Independent Cooperator, Germany
https://orcid.org/0000-0002-1507-2598

ABSTRACT: This report presents the methodological guidelines as well as some partial results of the study of referentiality in Mapudungun, a language isolate spoken in Chile and Argentina with different degrees of vitality. Reference in discourse encompasses two basic operations: the individuation of the referents and their anchorage in the discourse. Our research is part of a wider project which seeks to identify the structural resources used by the languages in referential operations. We investigate a set of structural and semantic parameters of referential expressions as they occur in a corpus of Mapudungun texts belonging to different genres. Some of the findings may represent general patterns of reference in natural texts, while others may be representative of specific Mapudungun genres. At a methodological level, our research shows that it is possible to substantiate hypotheses on reference and on discourse structure related to reference by hard figures, to characterize text genres by measurable semantic and structural properties, and to discover new phenomena which demand an explanation.
KEYWORDS: Mapudungun; Reference; Corpus-based analysis; Corpus annotation; Text genre

RESUMEN: Este artículo expone los lineamientos metodológicos y algunos resultados parciales del estudio de la referencialidad en mapudungun, una lengua aislada hablada en Chile y Argentina con distintos grados de vitalidad. La referencia en el discurso abarca dos operaciones básicas: la individuación de los referentes y su anclaje en el discurso. La presente investigación es parte de un proyecto más amplio que busca identificar los recursos estructurales utilizados por las lenguas en estas operaciones referenciales. Investigamos un conjunto de parámetros estructurales y semánticos de expresiones referenciales tal como ocurren en un corpus de textos en mapudungun pertenecientes a diferentes géneros. Algunos de los hallazgos pueden representar patrones generales de la referencia en textos naturales, otros pueden ser representativos de géneros mapuches específicos. A nivel metodológico, nuestra investigación muestra que es posible fundamentar hipótesis sobre la referencia y la estructura del discurso relacionada con la referencia cuantitativamente, caracterizar los géneros textuales mediante propiedades semánticas y estructurales medibles y descubrir nuevos fenómenos que requieren una explicación.
PALABRAS CLAVES: Mapudungun; Referencia; Análisis basado en corpus; Anotación de corpus; Género textual

1. Introduction

This research[1] is part of the Referentiality Project, a collaborative initiative, based both in European and South American institutions (Universität Erfurt and Universität Regensburg, Germany; Universidad de Buenos Aires and the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina; Universidad de Chile). It focuses on the study of the theory and typology of reference, with special emphasis on Amerindian languages. We aim to illustrate and analyze the methods used for the annotation and exploitation of texts, as well as some findings obtained from a Mapudungun text corpus.

We issue this report in a phase of our project where it is far from finished. What has been possible up to now is only a pilot study, based on a preliminary version of the theory (outlined in §2.1) and on a limited set of annotated data. The small size of the annotated corpus is mainly due to the demanding quality of the annotation. As a consequence, the corpus is too small to execute reliable statistics on it. We nevertheless approached some research questions with statistical methods just in order to try them out and to demonstrate what kind of research questions can be answered on the basis of our annotations.

The article is structured as follows: After this introduction (§1), in §2 we define some basic theoretical concepts and trace some typological features of Mapudungun relevant to this study. Section §3 presents the main lines of research on Referentiality in Amerindian languages. In §4, the structure of the corpus and the annotation system we use are explained. In §5, some concepts which are involved in the analysis are defined. §6 is the core section of the article. It encompasses the relevant working hypotheses and research questions tested in this article, as well as the preliminary answers. Finally, in §7 the main findings are summarized.

2. Background information

2.1 Referentiality

Referential operations are sensitive to the semantic categorization of referents. This takes the form of the **empathy hierarchy** (also known as animacy hierarchy) visualized in Diagram 1 (SAP = speech act participant):

[1]An initial version of this paper was presented at the 15/17 Symposium (Multiple methods for the study of indigenous languages) of the 56th Congreso Internacional de Americanistas, Salamanca, 15-20 July, 2018, coordinated by Liliana Sánchez (Rutgers University) and Marcus Maia (Universidade Federal de Rio de Janeiro). We are grateful to the coordinators and the participants of the Symposium, to two anonymous reviewers and to the journal editor for their comments. Part of this research was supported by the Volkswagenstiftung (Az. 86638). We wish to thank Johannes Helmbrecht and the Universität Regensburg for their support to this project and Anna Pamies for her cooperation in the research. Our recognition also goes to Deutsche Forschungsgemeinschaft (DFG) and the Argentine Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) for their support to the joint project titled "Referentiality in Two Southern Cone Languages: Mapudungun (isolate) and Wichi (Mataguayan)" (2014-2016), coordinated by Christian Lehmann and Lucía Golluscio. This project comprised Felipe Hasler's academic stay in Germany in 2015. Lastly, in 2018 Golluscio carried out a research stay at Universität Regensburg with a Georg Forster Research Award granted by Alexander von Humboldt Foundation. These favorable circumstances of international cooperation have enabled the authors to make progress in this research.

Diagram 1. Empathy hierarchy

Reference in discourse encompasses two basic operations: the individuation of the referents and their anchorage in the discourse. **Individuation** of a referent is conceived as a progressive narrowing. It starts from a concept, comprising a set of elements. Without any individuation, this affords generic reference. The first step in the individuation is non-specific reference to a subset of the elements constituting the set. At this step, interlocutors ignore the identity of these elements. The last step is specific reference, where the speaker has a set of individuals in mind that he can identify. **Anchorage** of a referent is a space where interlocutors find it. This may be encyclopedic world knowledge, shared experience, the speech situation or the universe of discourse. Once a referent has been individuated, it is present in the universe of discourse, and further mentions resume it from there. These operations manifest themselves in the structure of referential expressions, to be detailed in §5.

2.2 Mapudungun

Mapudungun (also called Mapuche or Araucanian) is a language isolate currently spoken with different degrees of vitality in central-southern Chile and southern Argentina. It is an agglutinative suffixing language with characteristics of polysynthetic head-marking typology. It exhibits a relatively free word order and a wide range of verbal valence-adjusting morphological categories. These include both valence-increasing (especially causatives and applicatives) (Golluscio 2007; Zúñiga 2010) and detransitivizing categories (passive voice, reflexive-reciprocals, and noun incorporation) (Salas 2006 [1992]; Baker et al. 2005; Evans et al. 2010). See examples (1-4) below.

(1a) *Lay ta ti kawell.*[2]

la-y	ta	ti	kawell.
die-IND.[3]	DET_1	DET_2	horse

'The horse died.'

[2] We have unified the transcription following the Mapuche Unified Alphabet developed by Robert Croese, Adalberto Salas and Gastón Sepúlveda in 1978 and adopted by the Linguistic Society of Chile (Sociedad Chilena de Lingüística 1988). Most orthographic symbols have roughly their expected value, with the following exceptions: *ü* high back (in stressed positions) or mid-central (in unstressed positions) unrounded vowel, *t̲* voiceless dental or interdental stop, *tr* alveopalatal retroflex affricate, *ch* palatal affricate, *d* voiceless interdental fricative, *l̲* voiced dental or interdental lateral, *l* voiced alveolar lateral, *ll* voiced palatal lateral, and *q* back unrounded glide. The representation is thus orthographic but includes morpheme boundaries and some sounds that are morphophonemically lost. The abbreviations used in the glosses are the following: $APPL_1$ 'applicative 1'; $APPL_2$ 'applicative 1'; $CAUS_1$ 'causative 1'; $CAUS_2$ 'causative 2'; DET_1 'determiner 1'; DET_2 'determiner 2'; DET_3 'determiner 3'; HAB 'habitual'; IND 'indicative'; INV 'inverse marker'; NMLZ 'nominalizer'; PL 'plural'; OBL 'oblique'; PASS 'passive'; PO 'primary object'; POSS 'possessive'; SG 'singular'; 1 'first person'; 2 'second person'; [3] 'third person, unmarked'.

(1b) *Lang**üm**i kawell*.
lang-**üm**-i kawell.
die-CAUS$_1$-IND.[3] horse
'He killed a horse.'

(2a) *Umawi ñi püchi che*.
umaw-i ñi püchi che.
sleep-IND.[3] 1SG.POSS small people
'My child fell asleep.'

(2b) *Umaw**el**n ñi püchi che.*
umaw-**el**-n ñi püchi che.
sleep-CAUS$_2$-IND.1SG 1SG.POSS small people
'(I) put my child to sleep'.

(3a) *Küdawün*.
küdaw-ün.
work-IND.1SG
'I worked.'

(3b) *Küdaw**el**fiñ Pedro*.
küdaw-**el**-fi-ñ Pedro.
work-CAUS$_2$/APPL$_1$-3.PO-IND.1SG Peter
'I made Peter work.' (or: 'I worked for Peter.')

(3c) *Küdaw**lel**fiñ Pedro*.
küdaw-**lel**-fi-ñ Pedro.
work-APPL$_2$-3.PO-IND.1SG Peter
'I worked for Peter.'

(4a) *Arelfiñ tañi waka*.
arel-fi-ñ ta=ñi waka.
lend-3.PO-IND.1SG DET$_1$=1SG.POSS cow
'I lent him my cow.'

(4b) *Arelngey tañi waka*.
arel-nge-y ta=ñi waka.
lend-PASS-IND.[3] DET$_1$=1SG.POSS cow
'My cow was lent.'

In contrast to the verb, nominal morphology is simple: Mapudungun lacks case marking; also gender and number are no grammatical categories in this language. The standard noun phrase construction is (DET +) MOD + N, DET being either an article (definite, indefinite), a demonstrative and/or a possessive pronoun, and a MOD being an attribute (adjective or noun), the latter being either in a "part-whole" or in a "possessor-possessed" construction. See the morphology of the following nominal constructions above: (1a) *ta ti kawell* 'the horse'; (2a-b) *ñi püchi che* 'my child'; (4a-b) *tañi waka* 'my cow'.

Particularly interesting for the study of referentiality is this language's complex system of personal reference with inversion (Salas 2006 [1992], Zúñiga 2006, Golluscio 2010). Mapudungun manifests an integrated inverse alignment system (Gildea 1994). This is governed by a version of the empathy hierarchy (Diagram 1) which combines the inherent

empathy associated with the status of each participant in the speech-act participant ranking with the discursive topicality related to a proximate/obviative opposition. See (5a-b).

(5a) *Pefiñ chi wentru*.
pe-fi-ñ chi wentru.
see-3.PO-IND.1SG DET_3 man
'I saw the man.'

(5b) *Peenew chi wentru*.
pe-e-n-ew chi wentru.
see-INV-IND.1SG-OBL DET_3 man
'The man saw me.' (Golluscio 2010: 715)

The alignment system in ditransitive constructions is governed by this hierarchy, too. It thus exhibits secundative alignment (Haspelmath 2005) in the encoding of ditransitive events; see (6). In both direct and inverse indexing, passivization, reflexive-reciprocals, and relativization, the recipient is aligned with the patient of monotransitive verbs (Golluscio 2010). See an example of passivization in (7).

(6) *Elufiñ kofke* [*tañi püñeñ*].
elu-fi_i-ñ kofke [ta=ñi püñeñ$_i$].
give-3.PO-IND.1SG bread DET_1=1SG.POSS child
'I gave bread to him [my child].'

(7) *Tukulelngekey ta ko* [*kako*].
tuku-lel-nge-ke-y-$Ø_i$ ta ko [kako$_i$].[3]
put-$APPL_2$-PASS-HAB-IND.[3]$_i$ DET_1 water mote
'Water is put [to the mote].' (Harmelink 1996: 253)

In non-local scenarios (i.e., the interaction between third persons), our research confirms the effect of the empathy hierarchy that determines the selection of voice, and, consequently, the prevalence of the semantic-pragmatic features of the referents over the thematic roles when determining the syntactic functions of the arguments in Mapudungun (Golluscio & Hasler 2017): The construction in (8) can only be used if the river is personified and assumes a certain level of agency that makes it rise in the hierarchy of empathy, otherwise the inverse construction (9) is used. The analysis of referential expressions in our text corpus will provide more evidence on this topic.

(8) *Chi lewfü yefiy Juana*.
chi lewfü ye-fi-y Juana.
DET_3 river carry-3.OP-IND.[3] Juana
'The river took Juana away.' (Golluscio y Hasler 2017: 83)

(9) *Juana yeeyew chi lewfü*.
Juana ye-e-y-ew chi lewfü.
Juana carry-INV-IND.[3]-OBL DET_3 river
'Juana was taken away by the river.' (Golluscio y Hasler 2017: 83)

[3] *Kako* 'mote'. In Chile, a "mote of wheat" refers to the grain of wheat that is boiled and peeled. The raw wheat is boiled with a mix of ashes from a native tree, which removes the husk of the wheat.

Some features of Mapudungun complex syntax are noteworthy: On the one hand, subordinate constructions are characterized by the use of nominalized verbal forms, with personal reference indexed on the possessive determiner preceding the non-finite verbal form. See ex. (10).

(10) *Kümey ta**mi** akun*.
küme-y ta=**mi** aku-n.
be.good-IND.[3] DET_1=2SG.POSS come-NMLZ
"[How] good that you have arrived." (Salas 2006 [1992]: 204)

On the other hand, desiderative constructions are monoclausal, being characterized by an inflected verb plus a desiderative particle (Hasler 2017; Smeets (2008: 175).

(11) ***Küpa** umawtuiñ*.
küpa umawtu-y-iñ.
want sleep-IND-1PL
'We want to sleep.' (Héctor Mariano, own data)

Finally, the Mapudungun language and culture exhibit a highly developed discursive theory and practice that distinguishes a rich scope of traditional oral genres and ways of speaking (Golluscio 2006), in which current western genres have been incorporated over the last decades.

3. Research on referentiality in Amerindian languages

The Referentiality Project is devoted to the area of reference and the linguistic means of achieving it in communication. The project seeks to identify the structural resources used by the languages in referential operations. Its empirical basis is constituted by corpora of diverse Amerindian languages, presently four South American languages — Mapudungun and Santiagueño Quichua (Andean), Wichi and Ayoreo (Chacoan), and one North American language, Hoocąk (Siouan). The texts are heavily annotated. The leading theoretical question is how reference works in natural texts of Amerindian languages. The methodological task is to test what kinds of research questions can usefully be put and answered by exploring this kind of corpus. The type of research question we are investigating may be illustrated by the following example set:

(1) Which proforms are used for different degrees of individuation, like specific and non-specific reference?
(2) Which factors determine the complexity of a referential expression? Possible factors include the semantic category of the referent, the distance between the present and the previous mention and the number of referents copresent in the context.
(3) Can a text genre be characterized by the prominence of referents of a certain kind, like human and specific referents or abstract and non-specific referents?

These and many more linguistic questions can be answered by analysis of the annotated corpora which we produce.

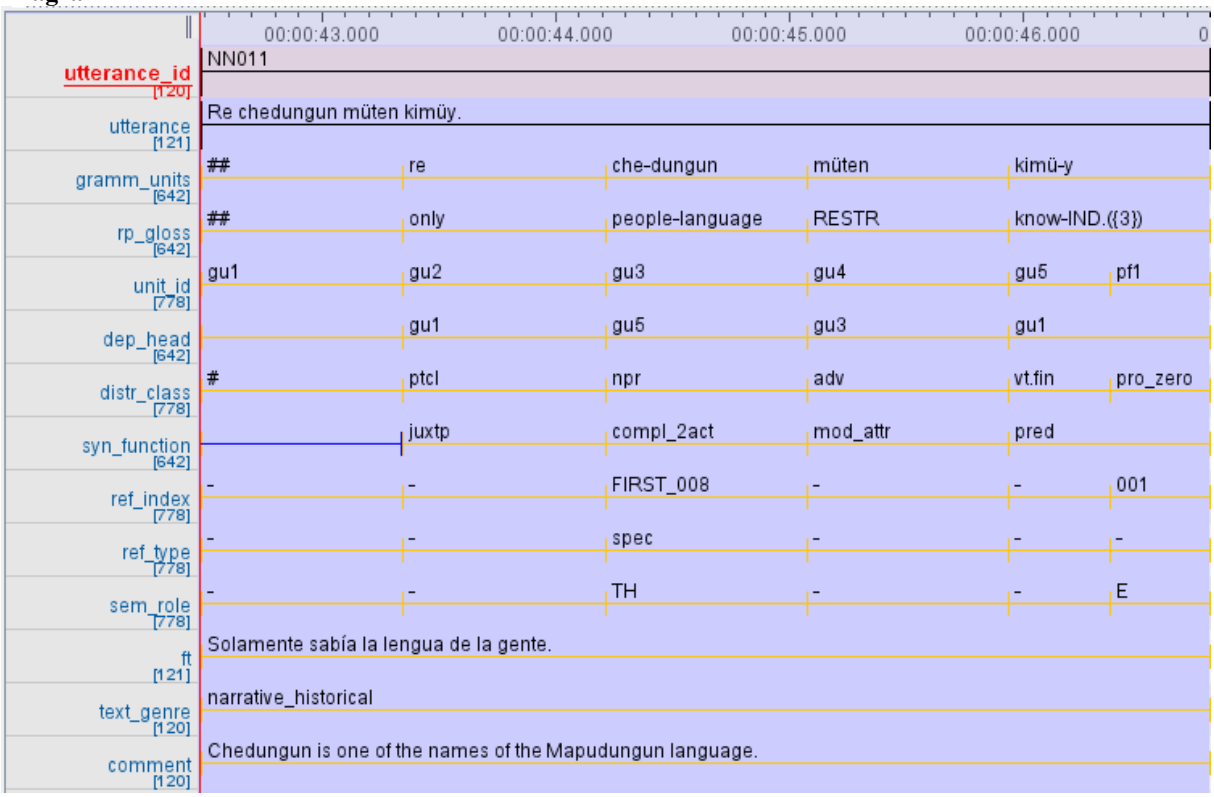
4. Corpus structure and annotation system

The texts of Amerindian languages represent different genres. Mythological and (auto-)biographic narratives are well represented; but there are also instructive texts and

dialogues. For this report we have chosen three texts belonging to different genres: two narratives and an instructional text. The first text is the *Sumpall*, which has been defined as a "mythological tale" (Salas 2006 [1992]: 219-221), a subclass of *epeo/epew* 'fictional tale'. The second narrative is the *Nawel Ngütram* ('Story of the Tiger') (Golluscio 2006: 175-181). The *ngütram* has been defined in the literature as a historical/non-fictional narrative. The instructional text is entitled *Chumngechi dewmangekey kako?* ('How to prepare *kako* 'mote'?') (Harmelink 1996: 253-254).

The texts have been transcribed and annotated in ELAN (The Language Archive, MPI for Psycholinguistics 2019). Diagram 2 shows one fully annotated sentence of the Mapudungun corpus.

Diagram 1. Screenshot of ELAN file



Horizontally, along the time line, the transcription is subdivided into utterances. On the vertical axis, the representation of these latter is arranged in tiers. For reference, each utterance gets a continual ID. The utterance is subdivided into a sequence of grammatical units, each of which receives an interlinear morphological gloss (Lehmann 2004). Inside each utterance, each grammatical unit has its ID. This is needed for the marking of dependency relations, done on the next lower tier, and for the marking of person forms, which may be referential expressions. On the next tier, each grammatical unit gets its distribution class assigned.

The distinct referents mentioned in a text are catalogued in a second file, called the referent list. There each referent has its ID. For instance, referent 008 is the people's language, and referent 001 is the protagonist of the whole story. In the ELAN file, all linguistic expressions which refer to one of these referents are marked for its ID. Thus, each referent index will occur one or more times throughout the text. In NN011, referent 008 is mentioned for the first time, while referent 001 has already been mentioned repeatedly in preceding utterances. On its first mention, every referent gets annotated for the type of reference – generic, non-specific and specific. The last tiers, finally, contain a free translation of the utterance, the textual genre of the passage in question and any free comment that may seem useful.

There are, thus, two files for every corpus text: an ELAN file which contains the running text with all the linguistic annotations at different levels, and a table which lists all the referents occurring in the former file with those of their linguistic properties which remain constant throughout the text. Both of these are XML files. The annotations are done manually, since no software is known that can delimit referential expressions, analyze them grammatically and ascertain all those semantic categories which are of interest to us; much less is there such software for Amerindian languages. The exploitation of all the information stored in the files, however, is done by JAVA classes which create a tree model of the XML structure and retrieve and combine the values found in the annotations.

5. Some analytic concepts

The empathy hierarchy of Diagram 1 is taken to define a rank order whose positions are assigned real values between 0.1 (bottom) and 1.0 (top). In the referent list, each entry is annotated for one of the categories of the hierarchy.

Moreover, referents are analyzed as individuated to different degrees. The degrees considered are listed in Table 1. They are assigned real values as shown.

Table 1. Individuation of referents

value	degree
0.25	generic
0.5	non-specific
1.0	specific

The grammatical category of a referential expression (second column of table 2) is constructed by means of the distribution classes and dependency relations as exemplified in the last column. Its weight is then read off the first column of Table 2.

Table 2. Weight scale of referential expressions

level	category	example
9	clause	that the child grew up in the jungle
8	(common) noun phrase with head and (clausal or adjectival) attribute	a child grown up in the jungle
7	(common) noun phrase with head and lexical possessive specifier	my friends' child
6	(common) noun phrase with head and determiner	this child
5	proper name	Susan
4	bare common noun	child
3.5	emphatic pronoun	French *moi* 'I'
3	neutral pronoun	she
2	clitic person form	French *je le vois* 'I see him/it'
1	bound person form (affix)	reads
0	ZERO person form.	∅

6. Research questions and preliminary answers

6.1 Empathy and individuation

Having thus operationalized the concepts of the position on the empathy hierarchy, the degree of individuation of a referent and the weight of a referential expression, we may assume as plausible that some traditional text genres can be characterized by the percentage of referents of different levels of the empathy hierarchy. This generates our first hypothesis:

Hypothesis 1 Empathy and text genre
The percentage of animate referents is significantly higher in narratives than in instructional texts.

We add a second hypothesis related to empathy:

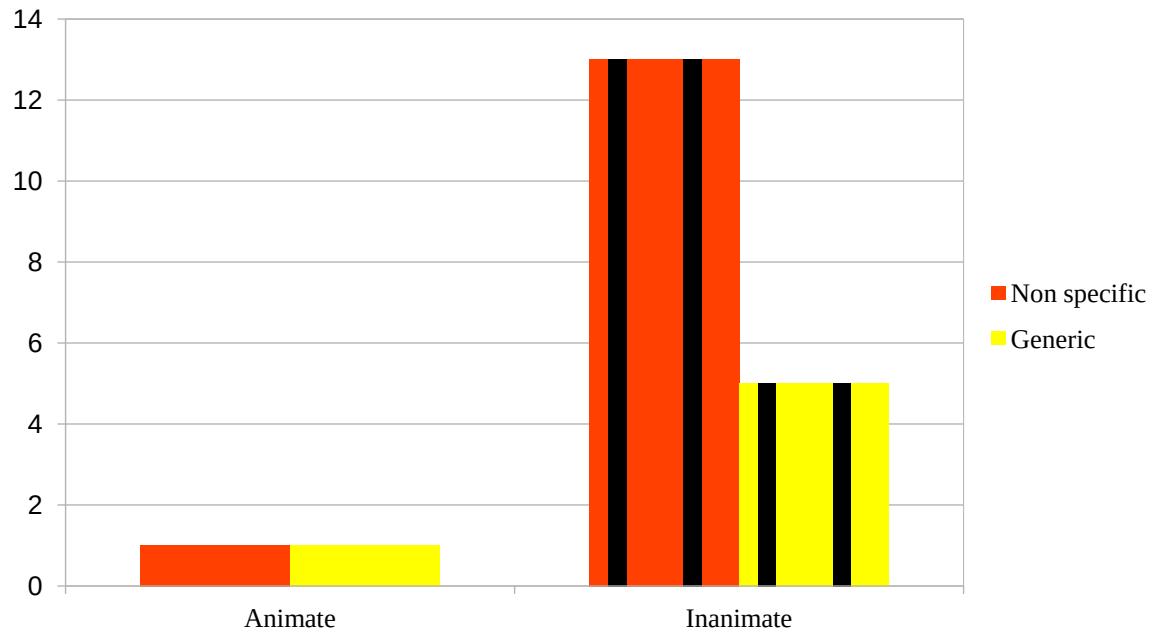
Hypothesis 2 Empathy and individuation
Elements higher up on the empathy hierarchy occur with relatively higher frequency with a highly individuated degree of referentiality.

We combine the data analysis for the two hypotheses. First, we analyze the relationship between empathy and individuation in the narrative texts, as shown in Table 3 and visualized in Diagram 3:

Table 3. Values of empathy and individuation in narrative texts

	specific	non-specific	generic	total
animate	26	3	1	30
inanimate	43	13	0	56
total	69	16	1	86

Diagram 3. Proportions of empathy and individuation in narrative texts



The important results concerning the narrative genre are the following:

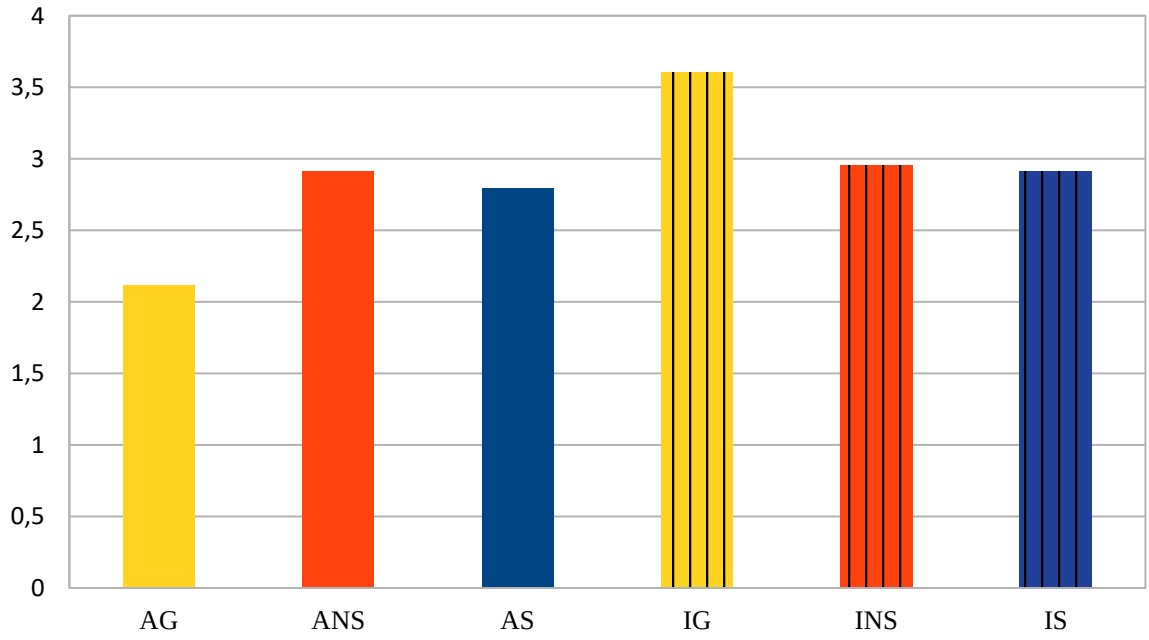
(1) 35% of all referents are animate.
(2) 80% of all referents are specific.
(3) The proportion of non-specific referents is higher for inanimate (23%) than for animate (10%) referents. I.o.w., animate referents are more typically specific.

For the instructional text, numerical relations between empathy and individuation are shown in Table 4 and visualized in Diagram 4.

Table 4. Values of empathy and individuation in the instructional text

	non-specific	generic	total
animate	1	1	2
inanimate	13	5	18
total	14	6	20

Diagram 4. Proportions of empathy and individuation in the instructional text



First of all, there are no specific referents in the instructional text; and instead, the percentage of generic referents (30%) is much higher than in narrative texts (1%). Given the low absolute number of occurrences, we abstain from drawing any particular conclusions concerning generic referents and instead subsume them under non-specific occurrences in what follows. A closer comparison of Diagram 3 and Diagram 4 allows us a first attempt at characterizing genres by the differences observed:

(1) While inanimate referents always outweigh animate referents, their percentage is significantly higher in instructional (90%) than in narrative (65%) texts. Hypothesis 1 is, thus, confirmed.
(2) While in narrative texts, most referents are specific, in instructional texts, all referents are non-specific or even generic. This applies equally to animate and inanimate referents.
(3) In narrative texts, animate referents have an even stronger tendency to be specific than inanimate referents. Conversely, non-specific referents tend to be inanimate.

Apart from differences between the genres, there is a generalization to be made over the entire set of texts:

- The average degree of individuation (according to Table 1) of animate referents is 0.89, while for inanimate referents it is 0.77. I.o.w., there is a positive correlation between the level of empathy and the level of individuation of a referent. Hypothesis 2 is, thus, confirmed.

6.2 Empathy, individuation and weight

Concerning the structural complexity of referential expressions for entities of different levels of the empathy hierarchy, we launch the following hypothesis:

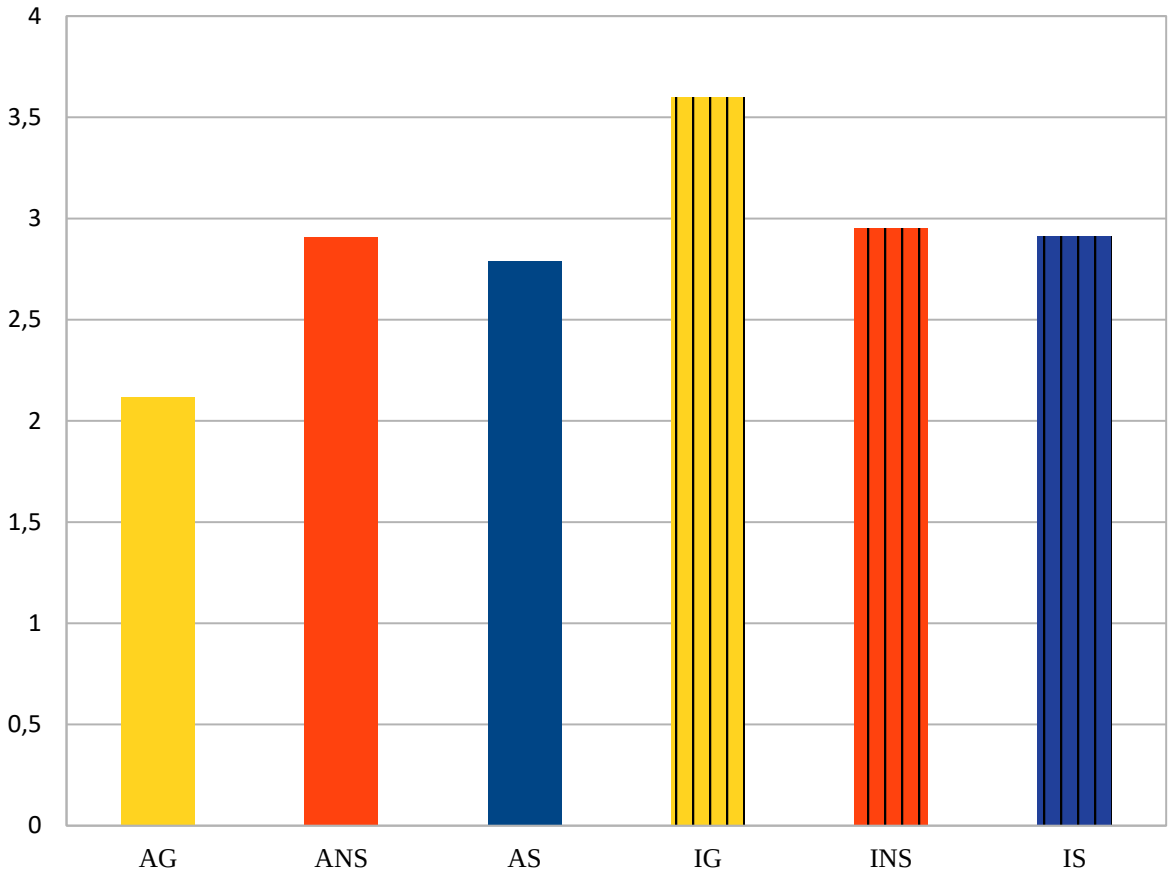
Hypothesis 3 Weight of referential expressions of different empathy and individuation
Referential expressions of higher levels of the empathy hierarchy require less grammatical apparatus for individuation; i.e., their referential expressions have relatively less weight.

For the relationship between empathy and individuation taken together, on the one hand, and the corresponding weight of referential expressions in the three texts, on the other, absolute figures are as shown in Table 5, they are visualized in Diagram 5.

Table 5. Weight of expressions of different categories in the three texts

category	weight
animate generic	2.165
animate non-specific	2.91
animate specific	2.79
inanimate generic	3.6
inanimate non-specific	2.95
inanimate specific	2.91

Diagram 5. Weight dependent on empathy and individuation in narrative texts



Legend: AG: animate generic; ANS: animate non-specific; AS: animate specific; IG: inanimate generic; INS: inanimate non-specific; IS: inanimate specific.

The first thing to be noted here is the low absolute number of human generic referents: The first column of Diagram 5 represents only two cases. If we disregard these as insufficiently representative, some generalizations suggest themselves:

(1) Animates tend to occur in referential expressions with less weight than inanimate ones, independently of both text genre and of degree of individuation.
In particular, and excepting the generic animates:
(2) At each of the individuation levels, animate referents have less weight than inanimate ones.
(3) At either of the empathy levels, specific referents have less weight than non-specific ones.

It may be concluded that the individuation of an inanimate referent requires weightier expressions than the individuation of an animate. We therefore regard Hypothesis 3 as confirmed.

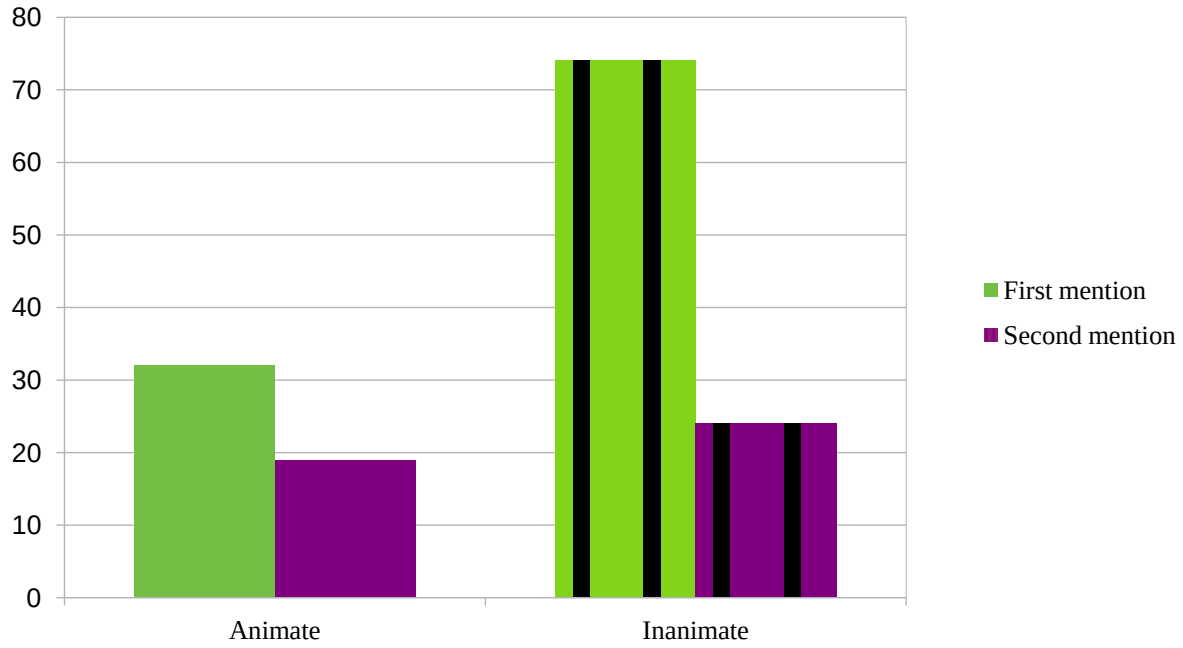
6.3 First and second mentions

Another question we investigated concerns first and second mentions of referents in a text. Given the annotations of the tier 'ref_index' of Diagram 2, the concepts of first and second mention of a referent may at first glance seem trivial. However, since all the kinds of constructions listed in Table 2 may constitute a referential expression, more than one of them may cooccur in a clause and share the same referent. There we stipulate that the weightiest of

these is the one to be considered. Assuming that animate referents have a higher chance of second mention in a text than inanimate referents, we then launch the following hypothesis:

Hypothesis 4 Dependence of second mentions on empathy
The proportion of second against first mentions in a text is higher for animate than for inanimate referents, independently of the text genre.

Diagram 6. Number of first and second mentions

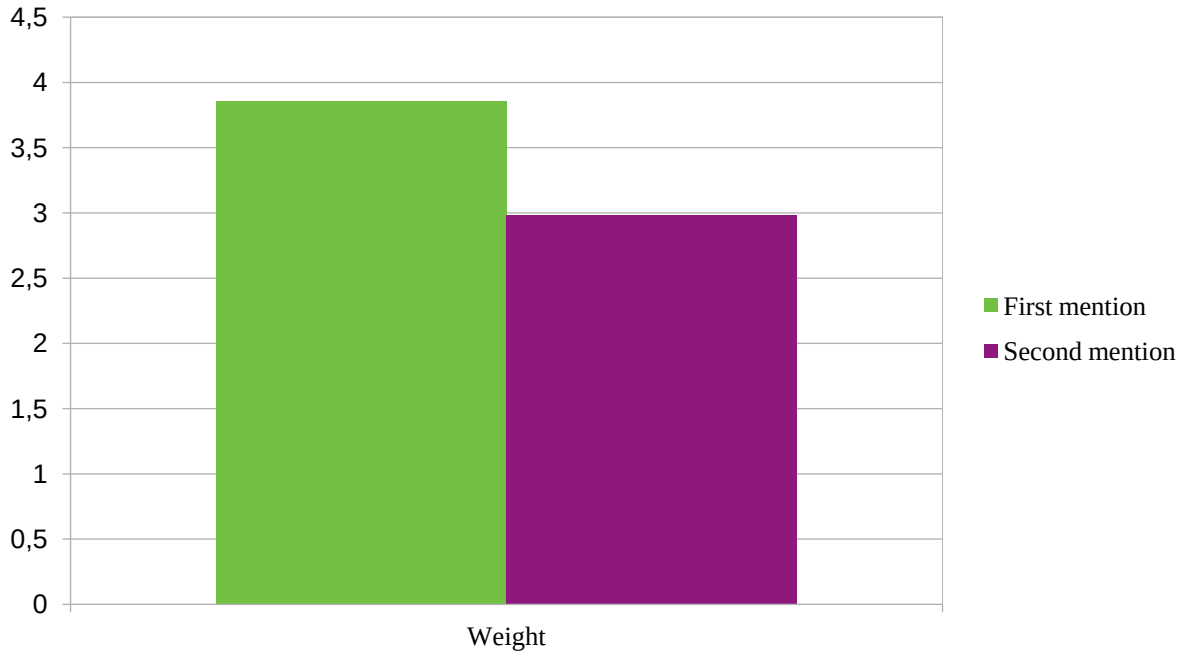


In Diagram 6, first mentions are taken from Table 3 and Table 4. Second mentions in the corpus are 19 for animate and 24 for inanimate referents. As may be seen, many referents are mentioned just once in a text, so the number of second mentions is lower than of first mentions for all kinds of referents. However, 59% of the animate referents are mentioned a second time, while only 32% of the inanimate referents are mentioned a second time. Thus, no matter what the proportion between animate and inanimate referents is in the text, in second mentions it is always changed in favor of animate referents. Hypothesis 4 is, thus, confirmed. This finding does not characterize any particular genre but instead indexes the relative persistence of human beings in human discourse. However, the phenomenon is even more pronounced in narratives. There, animate beings play a fundamental role in the creation of the narrative foreground, while most inanimate referents contribute to build location and time background; consequently, most of them need to be mentioned only once.

Next, we investigate the relative grammatical complexity of referential expressions, called “weight”, in first and second mentions, proposing the following hypothesis:

Hypothesis 5 Weight of first and second mentions
For any referent, the second mention will be less weighty than the first.

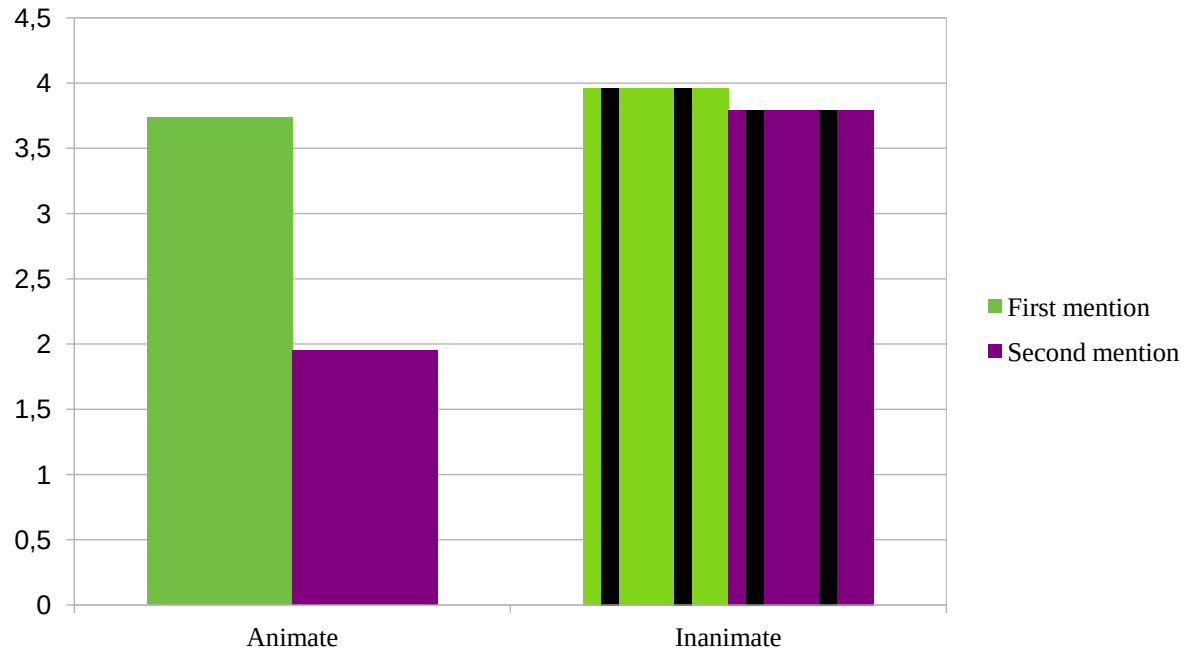
Diagram 7. Average weight of first and second mention



As is to be seen in the Diagram 7 second mentions, are on average, lighter than first mentions, so the hypothesis is confirmed: Introducing a new referent into the universe of discourse is, no doubt, a more costly operation than resuming one that is already present in the universe of discourse.

Next, we consider the average weight difference for first and second mentions of animate and inanimate referents. Here we have no hypothesis. Diagram 8 visualizes the findings.

Diagram 8. Average weight of animate and inanimate referents in first and second mentions



The decrease of the weight of second mentions as compared with first mentions is mainly a phenomenon of animate referents. For inanimate referents, the decrease does not seem statistically significant. This seems to show that not only the introduction, but also the resumptive reidentification is more costly for inanimate than for animate referents.

7. Conclusions

We have investigated a set of structural and semantic parameters of referential expressions as they occur in a corpus of Mapudungun texts. Some of the results may represent general patterns of reference in natural texts, others may be representative of specific genres of Mapudungun. The main findings are the following:

(1) The narrative genre is characterized by a significantly higher percentage of animate referents as opposed to the instructional genre.
(2) The narrative genre is characterized by a sizable percentage of specific referents, while no specific referents were found in the instructional text.
(3) It is more typical for animate than for inanimate referents to be specific. There is, thus, a positive correlation between a high level of empathy and a high degree of individuation.
(4) Consequently, the individuation of an inanimate referent requires weightier expressions than the individuation of an animate.
(5) Independently of the text genre, animate referents have a higher chance of second mention than inanimate referents.
(6) Second mentions are, on average, less weighty than first mentions. This, however, is more clearly true of animate than of inanimate referents.

At the methodological level, we concede that the manual annotation of texts of less-described languages for such linguistic parameters is extremely difficult and time-consuming. However, as we hope to have shown, the investment pays off. It becomes possible to substantiate hypotheses on reference and on discourse structure related to reference by hard figures, to characterize text genres by measurable semantic and structural properties and to discover new phenomena which demand an explanation.

References

Baker, Mark; Aranovich, Roberto; Golluscio, Lucía (2005). Two types of syntactic noun incorporation. Noun incorporation in Mapudungun and its typological implications. *Language* 81(1): 138-176. 10.1353/lan.2005.0003

Gildea, Spike (1994). Semantic and pragmatic inverse. 'Inverse alignment' and 'inverse voice' in Carib of Surinam. In T. Givón (ed.), *Voice and inversion* (Typological Studies in Language 28), pp. 187-230. Amsterdam/Philadelphia: John Benjamins.

Golluscio, Lucía (2006). *El pueblo mapuche. Poéticas de pertenencia y devenir.* Buenos Aires: BIBLOS.

Golluscio, Lucía (2007). Morphological causatives and split intransitivity in Mapudungun. *International Journal of American Linguistics* 73(2): 209-238. https://doi.org/10.1086/519058

Golluscio, Lucía. (2010). Ditransitives in Mapudungun. In Andrej Malchukov; Martin Haspelmath; Bernard Comrie (eds.), *Studies in ditransitive constructions. A comparative handbook*, pp. 711-756. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110220377.710

Golluscio, Lucía; Felipe Hasler (2017). Jerarquías referenciales y alineamiento inverso en mapudungun. *Revista RASAL* 1: 69-93.

Harmelink, Bryan M. (1996). *Manual de aprendizaje del idioma mapuche. Aspectos morfológicos y sintácticos*. Temuco: Ediciones Universidad de la Frontera.

Hasler, Felipe (2017). *Relaciones interclausales en mapudungun* (Doctoral dissertation). Buenos Aires: Universidad de Buenos Aires.

Haspelmath, Martin (2005). Argument marking in ditransitive alignment types. *Linguistic discovery* 3(1): 1-21. 10.1349/PS1.1537-0852.A.280

Lehmann, Christian (2004). Interlinear morphemic glossing. In Geert Booij; Joachim Mugdan; Stavros Skopeteas (eds.), *Morphology. An international handbook on inflection and word formation.* 2nd half-volume. (Handbücher zur Sprach- und Kommunikationswissenschaft, 17.2): 1834-1857. Berlin: W. DeGruyter.

ELAN - Linguistic Annotator version 5.8. *Manual*. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive https://archive.mpi.nl/tla/elan

Salas, Adalberto (2006 [1992]). *El mapuche o araucano: fonología, gramática y antología de cuentos*. Editado por Fernando Zúñiga. Santiago de Chile: Centro de Estudios Públicos.

Smeets, Ineke (2008 [1989]). *A grammar of Mapuche* (Volume 41 in the series Mouton Grammar Library). Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110211795

Sociedad Chilena de Lingüística (1988). *Alfabeto Mapuche Unificado.* Temuco: Universidad Católica.

Zúñiga, Fernando (2006). *Deixis and alignment. Inverse systems in indigenous languages of the Americas* (Typological Studies in Language 70). Amsterdam/Philadelphia: John Benjamins.

Zúñiga, Fernando (2010). Benefactive and malefactive applicativization in Mapudungun. In Fernando Zúñiga; Seppo Kittilä (eds.), *Benefactives and malefactives. Typological perspectives and case studies* (Typological Studies in Language 92), pp. 203-218. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.92.08zun

Recebido: 28/3/2020
Versão revista 1: 26/7/2021
Versão revista 2: 6/8/2021
Aceito: 8/8/2021
Publicado: 18/8/2021