ALBANIAN DEICTICS

A review of the meaning and functionality of demonstratives and deictic adverbs

Alexander Murzaku

TESI DI PERFEZIONAMENTO

Direttore
Pier Marco Bertinetto (Scuola Normale Superiore di Pisa)

Relatori Esterni
Paolo Di Giovine (Università degli Studi di Roma "La Sapienza")
Victor A. Friedman (University of Chicago)

Scuola Normale Superiore

Pisa, 2008

Contents

Prologue

> The extension, hardness, **impenetrability**, mobility, and *vis inertiæ* of the whole, result from the extension, hardness, **impenetrability**, mobility, and *vires inertiæ* of the parts; and thence we conclude the least particles of all bodies to be also all extended, and hard, and impenetrable, and moveable, and endowed with their proper *vires inertiæ*. [p. 161]
>
> …gravity does really exist, and act according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies, and of our sea. [p. 314]
>
> Isaac Newton
> *The Mathematical Principles of Natural Philosophy*
> Cambridge, 1686[1]

"…and of our words," I would dare to continue Newton's account of gravity. Gravity, one of the most mysterious phenomena of the universe, appears to influence words as well. Just like planets, certain words follow orbits around other words. But looking at the words themselves could be deceiving. The words investigated in this treatise, contain features and it is these features that can have the attribute of *impenetrability*. By observing the behavior of the words, we could find why they co-occur. Armed with a mathematical apparatus, every possible relation visible between words in the text will be dissected.

The thesis for this dissertation is that Albanian deictic words are defined through features which can be accessed through compositional analysis. Each feature can connect to a compatible feature in other words. It is these connections that make these words co-occur

[1] Translated in English by Andrew Motte in 1729. New Edition printed for H. D. Symonds in London 1803

(or gravitate around each-other.) These features can act by themselves or as part of a bundle of features. These bundles can be semantic or morphosyntactic. An example of an obvious morphosyntactic bundle is agreement, but we notice other non obvious feature connections. Is there some other kind of *meaning agreement* that makes certain words gravitate around each other? Is it possible to look at these fluid associations and observe repetitive connections which could even define the meaning of the collocated words?

The larger the amount of data, the more visible these meaning connections should become. A large corpus becomes an important need and a necessary first step… and having the most efficient mathematical apparatus for measuring connectivity between words is the second. While there is ample literature on how to measure collocations of words, there were no existing electronic corpora of Albanian language texts. The outcome of our work depended on the successful assembly of all the needed apparatus, data, and a good understanding of the category under investigation.

Some notes on Albanian

Albanian is an Indo-European language spoken in the contiguous territories covering Albania, Kosovo, parts of southern Montenegro, northern Greece and western Macedonia. It is also spoken by smaller numbers of ethnic Albanians in other parts of the Balkans, along the east coast and south of the Italian peninsula, in Sicily, in southern Greece, Ukraine and other countries where there are immigrant Albanian populations.

Figure 1: Europe, the Mediterranean, Balkans and Albania

Albanian territories in the Balkans remained for centuries as the theater of the conflict between Byzantine powers (Byzantium, Goths, Bulgars, Serbs and Ottomans) against the Roman Western powers (Rome, Venice, Normans and Anjou's), the first ones fighting to get access to the Adriatic Sea and the second ones trying to take control of the peninsula and from there Constantinople (Çabej, 1994:19). This central position contributed

in the definition and preservation of the Albanian ethnicity expressed in their way of life, religious history, traditions and folklore.

Origins: language and name

Albanian belongs to the family of the Indo-European Languages, along with the Indo-Iranian languages, Greek, the Latin languages, the Slavonic languages, the Germanic languages, etc. It constitutes a distinct branch in this family of languages and is not closely related to any of the modern Indo-European languages. The Indo-European origin of the Albanian language and the place it occupies in the family of Indo-European languages was determined and proved in the middle of the 19th century, following studies in comparative historical linguistics.

With the advent of significant computing power and sophisticated bioinformatics algorithms, the position of Albanian among Indo-European languages has been better determined as shown in Figure 2.

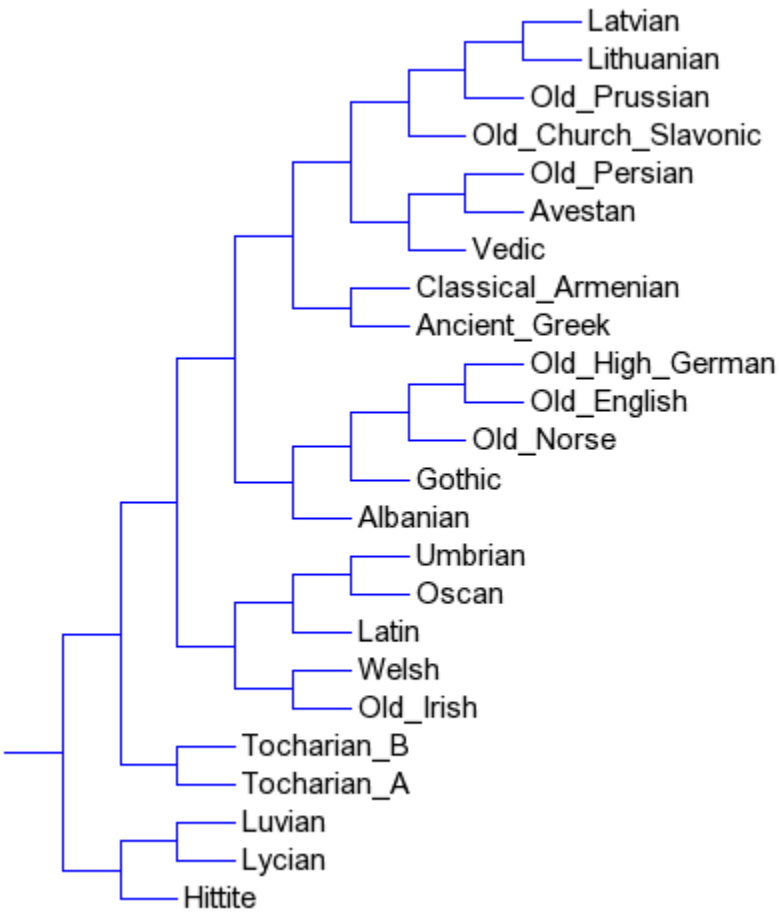


Figure 2: The Indo-European family of languages according to perfect phylogeny data analysis[2]

Following complex mathematical operations, Ringe et al. (2002) analyze the instantiation of several linguistic properties (also called characters) by twenty-four languages belonging to ten subfamilies of Indo-European. The tree representation constitutes the best

[2] A drawing of the Indo-European languages tree proposed in <http://www.cs.rice.edu/~nakhleh/CPHL/> as viewed on 10 March 10, 2008.

possible outcome from a computational method based on a *perfect phylogeny*[3] algorithm which analyses 370 characters of which 22 are phonological, 15 are morphological, and 333 are lexical. For example, if we consider the deictic particles of Hittite (proximal *k-* vs. distal *ap-*) and Albanian (proximal *k-* vs. distal *a-*), they could appear to be related to each other, therefore connecting the two languages together (see Table 1 and Table 2). In the phylogenic tree of Indo-European, this appears to be a case of homoplasy (features that emerge independently in more than one branch. See also footnote 3).

here		there		over there	
këtu	kā	aty	apiya	atje	apēda

Table 1: Albanian and Hittite deictic adverbs

this		that	
ky / kjo	kī (kē)	ai / ajo	apāt

Table 2: Albanian and Hittite deictic pronouns

Following the same algorithms, Nakhleh et al. (2005) establish the creation of a distinct Albanian branch sometime between 3000 and 2000 BCE as represented in Figure 3.

[3] Phylogenic tree represents an evolutionary process, where new species split off from existing ones, thus creating the diversity of life forms or languages we know today. A key issue in evolutionary biology/linguistics is to reconstruct the history of these events. Given the properties of the leaf nodes, reconstruct what the tree is. Sometime features emerge independently in more than one branch. This is called homoplasy and is generally inescapable in real data. Homoplasy is a poor indicator of evolutionary relationships because similarity does not reflect shared ancestry. Sets of characters that admit phylogenies without homoplasy are said to be compatible. Phylogenies that avoid homoplasy are called perfect and the character compatibility problem is called the perfect phylogeny problem.

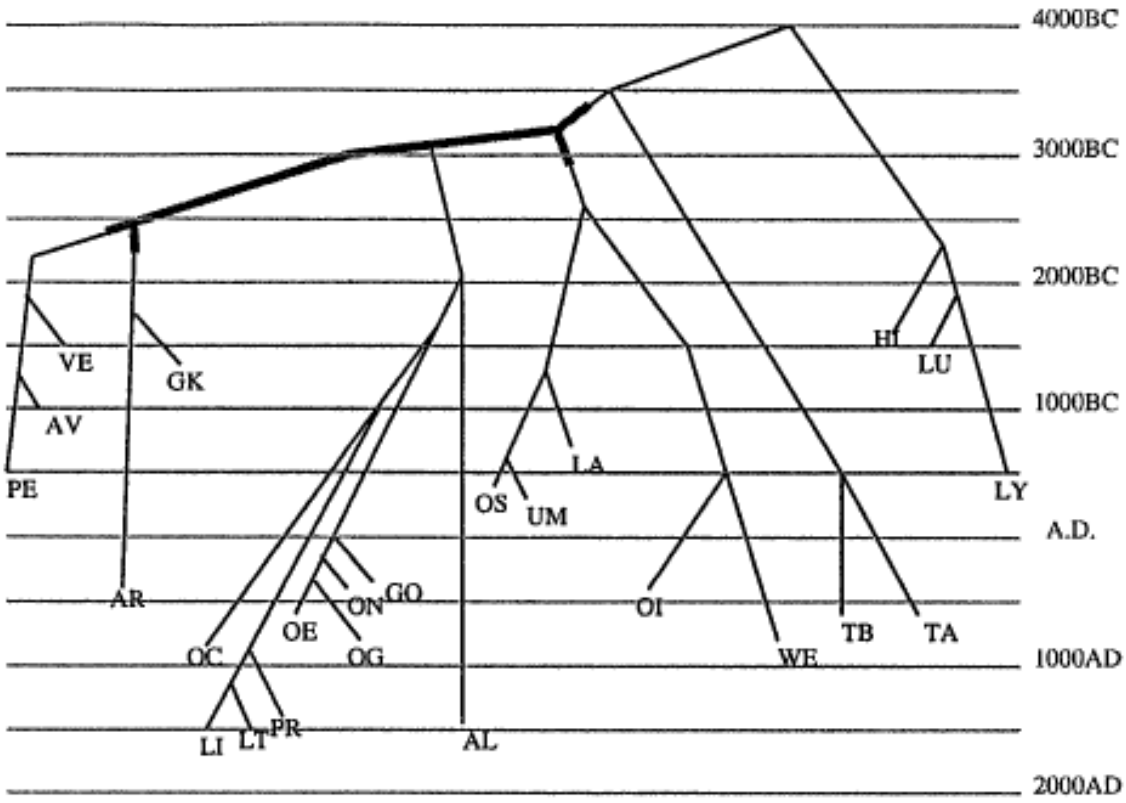


Figure 3: One of the possible trees[4] based on perfect phylogeny

The origin of Albanian language is one of the most disputed issues in the field of Albanology. Çabej, in his *History of Albanian Language* (1976c), argues that the roots of the Albanian language are to be found in one of the ancient languages of the Balkan Peninsula. Two main theories have circulated in linguistic literature with regard to the Albanian language: one which finds the roots of Albanian in the Illyrian language and another one which finds them in the Thracian language. As Victor Friedman (1997) observes, the debate over Thracian versus Illyrian provenance has its basis in serious evidence, and in any case

[4] Since Albanian has lost nearly all of the diagnostic inflectional morphology (as well as a large proportion of inherited words on the IE wordlist), it can attach anywhere in the thick lines in the tree. (Nakhleh et al. 2005)

there is no rival group claiming descent from either of these two languages. However, the logic of ideologically based contestation would argue that if Albanian is of Thracian origin then the Albanians arrived in northern Albania at around the same time as the Slavs and in southern Albania after the Greeks, and that therefore their current claims to sovereignty are somehow less legitimate, whereas if they are descended of the Illyrians then their claims are somehow more legitimate. According to Çabej, the Illyrian theory has had a broader historical and linguistic support. Some of these arguments that support the Illyrian origin of Albanians and their language are:

1. Albanians are currently living in some of the territories, which were inhabited by Illyrians in ancient times; on the other hand, historical sources do not speak of any Albanian migration from other territories to the present ones.
2. A number of linguistic elements such as names of things, tribes, people, etc., of Illyrian origin, are explained in the Albanian language.
3. The ancient toponymic forms from the Illyrian territories, as compared to the corresponding present-day forms, prove that they have evolved following patterns established by the phonetic rules of Albanian.
4. Relationships between the Albanian language and the ancient Greek and Latin suggest that the Albanian language took shape and developed side by side with these two neighboring languages on the shores of the Adriatic and Ionian seas.
5. Archaeological findings and cultural heritage (myths, music, costumes, etc.) testify to the cultural continuity from the ancient Illyrians to the present-day Albanians.

The name of the country Albania in Albanian is derived from the Albanian word for Albanian: Albanians call the language *Shqip*, the country *Shqipëri* and themselves *Shqiptar*. The general name of Albania traditionally referred to a restricted area in central Albania where

the tribe of the Albanoi and the city of Albanopolis were situated.[5] While non-Albanians fluctuated between the form in *alb-* and in *arb-* (by both Byzantines and Latins), Albanians always have preferred the form in *arb-*. The form in *arb-* is preserved among both Tosks and Ghegs.[6] The origin of the current official names of the language and country *Shqip* and *Shqipëri*, which may well be derived from the word *shqiptoj*[7] meaning "pronounce clearly, intelligibly," are still disputed. Other theories connect this word to the eagle (*shqipe* < *shqype* in Albanian)[8]. Çabej (1994:15) cites also documents from 1368-1402 that show family names such as Schibudar, Schebudar, Schabudar, and Scapuder. However, its usage as the name of the language is first attested in 1665 in the introduction of Cuneus Profetarum where Pjetër Bogdani writes *Gramatike Latìn e Scqip*.

First written records and the alphabets

Even though Albanian is accepted to be one of the ancient languages of the Balkans, its written records date back only to the 15th century. In a document written in Latin, the archbishop of Durrës reports of his recommendations given to the populace. In the case of

---

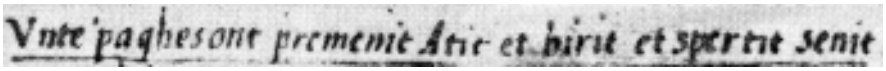
[5] Ptolemy (2nd century CE) included in his world map, just east of Lissus, the tribe of the Αλβανοί and the city of Αλβανόπολις

[6] The word used among Albanians in Italy and in Greece is *Arbëresh* which would attest to its existence in Tosk until late Medieval times. The word used among the Albanians of Borgo Erizzo in Dalmatia is *Arbënesh* attesting to its usage in Gheg until the 17th century. As for the Turkish word, Çabej assigns to it a clear Greek origin (Αρβανίτης/Αρναβίτης > Arnaut)

[7] Proponents of this theory are Meyer, Jireček, and Weigand (Çabej, 1994).

[8] Lambertz, Durham and Çabej show that there are many cultures and people that identify themselves with birds. Historic and ethnographic facts about the eagle being like a totem among Albanians would support this theory.

illnesses, when there were no priests nearby, parents could baptize their children themselves by pronouncing the baptismal formula recorded in the Gheg dialect:



Unte paghesont	premenit	Atit et	birit et	spertit senit
I baptize thee	in the name of	the Father, and	the Son, and	the Holy Spirit

Edith Durham (1909/2000:4-5) quotes a certain Frère Brochard in 1332 who writes about the Abbanois, living in the tribal districts of Upper and Lower Pulati, the diocese of Sappa and the diocese of Durazzo, and having a language quite other than Latin but using Latin letters in their books. Another Albanian text dating back to early 16th century was found inside a Greek manuscript. It contains extracts translated to Tosk (southern Albanian) from the Gospel according to Matthew and is written in the Greek alphabet.

These monuments and witnesses show that from the early stages of its written form, the Albanian language is proved to be written in two dialects, in the northern dialect (Gheg) and in the southern dialect (Tosk), and in two alphabets, the Latin and the Greek depending on whether the corresponding territory was under Venetian or Byzantine influence. The history of the alphabets used by Albanians follows the historical periods of Albania itself. Indeed, later on, Albanian has also been written in Arabic script which was enforced by the Sublime Porte. As the fall of the Ottoman Empire was approaching, a fatwa was sent to Muslim clergy in Albanian territories ruling that the Latin script was not in accordance with Islamic Law and thus prohibiting its use in schools. However, beside the writings of some poets of the 18th century, only one Albanian newspaper ever appeared in the Arabic script, and it lasted for a brief period (Gawrych, 2006). To a lesser extent, Cyrillic too was used. In 1875, Gjorgji Pulevski, who was from western Macedonia, published *Three Language Dictionary* (Macedonian, Albanian and Turkish) in a modified Cyrillic alphabet. This phonetic alphabet

allowed the accurate transcription of words and parallel texts in the three concerned languages (Kramer, 2001).[9] It appears that the choices of graphic systems in Albania have always corresponded with ideological/political orientations. Finally, Albanians adopted the Latin script. As Alberto Mioni (2000) points out, adoption of Latin script usually indicates an acceptance of a modernizing and internationalist ideology. Standard Albanian is written in the Latin alphabet which was one of the two agreed upon in 1908 at the Congress of Monastir (Bitola, Macedonia). Aiming to be a phonetic alphabet, it consists of 36 letters. 25 of these letters are simple (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, x, y, z), 9 are digraphs (dh, gj, ll, nj, rr, sh, th, xh, zh) and 2 diacritic letters (ë, ç).

The study of Albanian

Albanological studies began in the early 19th century with scholars such as Hahn and Pouqueville and even earlier when Leibnitz suggested that it would be a good idea in Albanian to distinguish between what is native to the language and what has been borrowed.[10]

Following scientific criteria and methodologies, Franz Bopp established the foundation of the linguistic classification of Albanian, its grammatical forms as well as the

---

[9] The use of phonetic alphabets by Albanians and Macedonians for the transcription of Turkish was of significant linguistic interest since they included vowels as opposed to the vowel-less Arabic script. These efforts played a significant role in the future transition of Turkish itself being written with the Latin alphabet (Trix, 1999).

[10] Gottfried Wilhelm Leibniz, *Opera Philologica* Hildesheim: Georg Olms, 1989. [Hanover, 10 December 1709]. Translated from French by Robert Elsie. In <http://www.albanianhistory.net/texts/AH1705.html>. Accessed on 6 March 2008.

Indo-European character of many words.[11] During this period, the most significant contribution to the etymological study of Albanian is attributed to Gustav Meyer with his Etymologisches Wörterbuch der albanesischen Sprache of 1891 followed by Albanesische Studien (I-IV) during 1883-1897, Neugriechische Studien (I-IV) during 1894-1895, and Türkische Studien (I) in 1893. He looked at the lexical data of Albanian in their entirety comparing it to both the Indo-European stock as well as to the possible borrowings during history. His legacy was followed by Holger Pedersen who found phonetic rules and morphological structures which opened a new way into the comparative etymological studies. Norbert Jokl looked at the origin of words as well as the relations between the Balkan languages from a different point of view: he used extensively Albanian material, both written and oral, he compared ethnographic and cultural data providing a more holistic approach to the solution of various linguistic problems as he argued in his Linguistisch-kulturhistorische Untersuchungen aus dem Bereiche des Albanischen published in 1923. During the same period, Albanian and its relations with the other Balkan languages and cultures were the object of studies by Johann E. Thunmann (1746-1778), Franz Miklosich (1813-1891), Wilhelm Meyer-Lübke (1861-1936), Gustav Weigand (1860-1930), Carlo Tagliavini (1903-1982), Stuart Mann (1905-1986), and others. Among Albanians there were significant contribution by Demetrio Camarda in Italy, Kostandin Kristoforidhi and Aleksander Xhuvani in Albania. Eqrem Çabej (1908-1980) was undoubtedly the most outstanding modern Albanian scholar in the field with a voluminous work ranging from etymology and linguistics to folklore and literature, and, not unlike his professor, Jokl, giving

[11] Franz Bopp. 1854, Über das Albanesische in seinen verwandtschaftlichen Beziehungen. Berlin.

the field of Albanian etymology an interdisciplinary focus. As Rexhep Qosja points out[12], this variety of interests has determined the diversity of themes, and volume of his creative work, as well. Beside this tradition in Albanian linguistics, today, the field is enriched by studies in many areas following various modern linguistics theories. In addition to Balkan countries' scholarly activity, there are also many centers around the world including several universities in Italy, the U.S.A., Russia and Germany.

[12] Eqrem Çabej – Dijetari i madh. In *Gjurmime Albanologjike, 36*. 2006. Prishtinë, Kosovë: Instituti Albanologjik.

Dialects

Figure 4: Approximate line of demarcation between Gheg in the North and Tosk in the South

The two principal dialects, Gheg in the north and Tosk in the south, through a significant bundle of isoglosses are divided by the course of the river Shkumbin in central

Albania and then along the river Drin through the middle of Struga in the Republic of Macedonia (Friedman 2004). These two dialects have been diverging for at least a millennium, and their less extreme forms are mutually intelligible. Southern Tosk dialects spoken in the albanophone colonies in Greece and Italy are close to the Gheg of the 16th and 17th centuries indicating a unity of the Old Albanian which lasted until the late medieval period (Çabej, 1994:20).

Phonologic systems

The phonologic differences in the vowel system consist mainly of the presence of nasality and length. The vowel system goes from five (/a/, /e/, /i/, /o/, /u/) in Gjirokastër to nineteen (long, short and nasal) in a few Gheg dialects where the short and the long vowels are not merely allophones as in most other dialects

The system of consonants includes /p/, /b/, /t/, /d/, /c/ (written as q), /ɟ/ (written as gj), /k/, /g/, /ʦ/ (written as c), /ʣ/ written as x, /ʧ/ (written as ç), /ʤ/ (written as xh), /θ/ (written as th), /ð/ (written as dh), /f/, /v/, /s/, /z/, /ʃ/ (written as sh), /ʒ/ (written as zh), /h/, /m/, /n/, /ɲ/ (written as nj), /j/, /l/, /ɫ/ (written as ll), /r/ (written as rr), and /ɾ/ (written as r). However, in some Gheg dialects /c/ (written as q) and /ʧ/ (written as ç) on one hand and /ɟ/ (written as gj) and /ʤ/ (written as xh) are merged into one or the other consonant of the pair.

Other differences are the diphthong /ua/ in Tosk which has the equivalent /ue/ in Gheg (*grua/grue* 'woman'), the initial sequence /va/ in Tosk has the equivalent /vo/ in Gheg (*vatër/votër* 'hearth'). Tosk is characterized by rhotacism (the changing of /n/ to /ɾ/ (*rān/rërë*

'sand'); in Tosk the consonant clusters /mb/, /nd/, etc. are retained whereas in Gheg they are simplified to /m/, /n/, (*mbush/mush* 'to fill', *vend/ven* 'place/country').

The phonological system of standard Albanian includes 7 vowels and 29 consonants. The set of seven vowels is created by not using length and nasality and it includes the five vowels (/a/, /e/, /i/, /o/, /u/) present in every dialect, /ə/ (written as ë) more prominent in Tosk, and /y/ which is not present in the dialect of Gjirokastër (Tosk) and in some dialects of Dibër (Gheg).

Morpho-syntax

Albanian is structurally an analytical-synthetic language, with a dominance of synthetic elements tending towards being analytical. Some of its phonetic and grammatical features date back to the Indo-European period, others have developed later. Word order is generally free but the most common form is SVO.

The most visible morpho-syntactic differences between Gheg and Tosk are the presence of infinitive in Gheg, e.g. *me shkue* ('with' + short participle of 'go') and the formations of future, e.g. 'I will go' becomes *kam me shkue* ('have' + infinitive) in Gheg as opposed to the Tosk *do të shkoj* (particle based on 'want' + present subjunctive verb).

Tripartite Indo-European gender is reduced to two in modern Albanian. Beside masculine and feminine, there are still remnants of the neuter in some deverbal nouns like *të folurit* 'act of speech,' some dialectal mass nouns like *ujtë* 'water' and *vajtë* 'oil,' and the article *të* found in these forms.

Albanian nouns, pronouns, and adjectives inflect according to definiteness, number, gender, and case. Verbs are conjugated according to person, number, tense, aspect, voice, and mood.

As in many other languages, the definite article has a clear pronominal origin. In modern Albanian, as elsewhere in the Balkan Sprachbund, it is affixed at the end of the word and has merged its functionality with that of marking number, gender and case.

The indefinite article derives from the numeral 'one' *një* and does not have gender. There have been unsuccessful proposals to include the Gheg form *nji* for feminine in analogy to the numeral 'three' which is *tri* for feminine and *tre* for masculine.

Case endings have Indo-European roots and have followed similar paths with other languages. Albanian language has a generally fixed stress during inflection. In most cases, especially in the noun system, the stress falls on the last syllable of the stem. The three most frequent cases are nominative, accusative and dative which are mostly found in the syntactic roles of subject, direct object and indirect object respectively. Ablative has its own ending (-*sh*) only in indefinite plural. Genitive is formed by preposing the dative form with one of the particles of concord *i*, *e*, *të*, and *së*, usually translated in English by the preposition 'of.'

The appropriate particle of concord, appearing also in front of some adjectives, pronouns and nouns (usually deverbal), is normally chosen in agreement with the gender, number and case of the preceding noun. As for the adjectives, it should be noted that there are also adjectives that are not preceded by a particle of concord, such as *trim* 'courageous,' *besnik* 'faithful,' etc. Both these adjectival constructs are thought to be inherited from Proto-Albanian (Demiraj, 2002). Below are some usage examples of the particle of concord.

libri 'book' [m, sg, NOM]	*i* 'of' [m, sg, NOM]	+	*djalit* 'boy' [DAT/GEN/ABL] *kuq* 'red' [m] *atij* 'that one' [DAT/GEN/ABL]
fletoren 'notebook' [f, sg, ACC]	*e* 'of' [f, sg, ACC]	+	*vajzës* 'girl' [DAT/GEN/ABL] *kuqe* 'red' [f] *tij* 'that one / his' [DAT/GEN/ABL]
librave 'book' [m, pl, DAT/GEN/ABL]	*të* 'of' [m, pl, DAT/GEN/ABL]	+	*vajzës* 'girl' [DAT/GEN/ABL] *kuq* 'red' [m] *saj* 'that one / her' [DAT/GEN/ABL]
fletores 'notebook' [f, sg, DAT/GEN/ABL]	*së* 'of' [f, sg, DAT/GEN/ABL]	+	*djalit* 'boy' [DAT/GEN/ABL] *kuqe* 'red' [f] *këtij* 'this one' [DAT/GEN/ABL]

Table 3: Some usage examples of the particle of concord

A similar pattern is followed by the usage of clitics in Albanian better know as clitic doubling. Doubling is mandatory when verbs are followed by an indirect object (in DAT). When verbs are followed by a direct object (ACC), clitic doubling becomes a syntactic marker of topichood. Direct object clitic doubling in Albanian produces information structure in a systematic way: doubled DPs are unambiguously interpreted as topics (Kallulli, 2001).

The complete list of clitics is *më* [1st, sg, DAT/ACC], *na* [1st, pl, DAT/ACC], *të* [2nd, sg, DAT/ACC], *ju* [2nd, pl, DAT/ACC], *i* [3rd, sg, DAT], *u* [3rd, pl, DAT], *e* [3rd, sg, ACC], and *i* [3rd, pl, ACC]. For verbs that take both direct and indirect objects, clitics can be contracted together, as in the table below:

DAT	ACC (Sg/Pl)	
më	e	ma
	i	m'i
të	e	ta
	i	t'i
i	e	ia
	i	ia
na	e	*na e*
	i	*na i*
ju	e	jua
	i	jua
u	e	ua
	i	ua

Table 4: Clitic contractions

In the verbal system, beside person and number, endings mark the mood of the inflected verb as well e.g. *të lash* 'wash [2nd, sg, subjunctive].' Medio-passive is formed by its own set of endings. Besides indicative, Albanian has imperative, subjunctive, conditional, admirative and optative moods. Similarly to Italian (Bertinetto 1986:19), the verb can be in Present, Imperfect, Simple Perfect, Compound Perfect, Pluperfect, Pluperfect II, Simple Future, and Compound Future. Compound tenses are formed by using the auxiliary *kam* 'to have' followed by the past participle. The verb *jam* 'to be' is used to form the compound tenses of medio-passive forms. Medio-passive's Simple Past is formed by the clitic *u* + stem. E.g. the verb *laj* 'to wash' forms its 3rd singular Present by adding to the stem *la-* the ending *-n* as in *la-n*. Simple Past is *la-u*, Compound Past is *ka la-rë*. The Present medio-passive *la-h-et*, forms Simple Past *u la-ø* and Compound Past *është la-rë*.

Lexicon

From the lexical point of view, since Albanian has coexisted side by side with the Greek and Latin worlds, it is natural that it has borrowed many words from these languages. Phonetically, these words have been affected through usage of Albanian speakers. For example, we find in Studime Etimologjike në Fushë të Shqipes (Çabej, 1976a and 1976b) the

derivation from Old Greek *χυλός* (most probably Doric where '*υ*' is pronounced /u/) which makes it to today's Albanian as *qull* 'grits,' or from Latin *factura* that gives us *fytyrë*, etc. Language contacts through centuries have produced many other borrowings from Latin and Slavonic languages, as well as Turkish. Modern Albanian, through the adoption of technologies, processes, legal and economic concepts, has been flooded especially by Italian, English, and French words. Despite the numerous borrowings during centuries, Albanian has retained its originality as a separate Indo-European language. Words considered of Albanian or Pre-Indoeuropean stock are, for example, krimb 'worm,' kërmill 'snail,' ngrys 'becomes dark,' nis 'begin,' zi 'black,' zog 'bird,' etc. as discussed in Çabej (1976a, 1976b).

Standard Albanian

Following a long process started in 19th century called the National Renaissance, between 1968 and 1972 a Tosk based standard was decided at what came to be known as *Kongresi i Drejtshkrimit* 'the Congress of Orthography.' This standard was promulgated in Albania and adopted by the ethnic Albanians of former Yugoslavia. As Victor Friedman (2004) observes, "the lack of standardization before the publication of the standard reference tools of the 1970's made it extremely difficult for the foreign learner to know which forms to memorize." I should add that even native Albanians do not know which forms are considered more correct and use variants interchangeably. The confusion between *pritni* or *prisni* 'wait [2nd pl indicative present],' *u gjet* or *u gjend* 'find [3rd sg medio-passive aorist],' the past participles of *rri* 'sit' which could be either *ndenjur* or *ndejtur* or *ndënjur* is present even in our corpus where *pritni* occurs 19 times while *prisni* 84 times, *gjet* occurs 668 times while *gjend* 194, and *ndenjur*, *ndejtur* and *ndënjur* occur respectively 196, 9 and 1 times.

Language change

In a paper presented in the 1998 meeting of the Società Italiana di Glottologia, Alberto Mioni (2000) proposes a taxonomy of the macrocauses of language change. They include natural catastrophes which cause massive migration, population, depopulation, and repopulation. There are also sociopolitical and economic catastrophes such as wars, invasions, organization of states, creation and fall of empires, nationalism, regionalism, religion, and revolutions.

The effects of these macrocauses in the change of language and its usage include changes of the standard or in the standard, discontinuity of the educational system, graphic system choices and relations with other languages. Inside the language, the corresponding triggering event that leads to extensive systematic change is the insertion or removal of a category from a subsystem (such as phonology or lexicon). Labov (2007) labels these changes when languages are in contact as transmission and diffusion. "The differences between the two are absolute: one copies everything; the other is limited to the most superficial aspects of language: words and sounds. The contrast between the transmission of change within languages and diffusion of change across languages is the result of two different kinds of language learning. On the one hand, transmission is the product of the acquisition of language by young children. On the other hand, the limitations on diffusion are the result of the fact that most language contact is largely between and among adults." (Labov, 2007:349)

During the 20th century, to one extent or another, in a compressed timeline, Albania went through most of these catastrophes and their effects on language. We will focus on the last of these changes that involve contemporary Albanian.

The dramatic collapse of communist rule in Eastern Europe in 1989 had a devastating effect on the internal social and political situation in Albania as well. Massive demonstrations against the communist rule followed by liberalization and democratization in Eastern Europe began to affect Albania in 1990. After the unrest in various parts of the country, the students' protests and massive popular demonstrations, on December 31, 1991, was drafted an interim constitution intended to replace the constitution of 1976. The draft completely omitted mention of the Albanian Labor Party. It introduced a system with features similar to those of a parliamentary democracy. A multiparty system was introduced in time for the general elections of Spring 1991. The economic and political crisis in Albania, and the ensuing breakdown of public order then and again in 1997, plunged the educational system and cultural life into chaos. The highly structured and controlled educational environment that the communist regime had painstakingly cultivated in the course of more than forty-six years was abruptly shattered. While Albania found itself quickly with at least 25 different newspapers[13], 46 FM radio stations, 65 TV stations[14] being published or supported by political parties, various institutions, and private entities, there were at the same time 2,500 schools that were ransacked and at least 2,000 teachers that emigrated abroad by 1992. UNESCO's 26 April - 7 May 2004 mission report states that Albanian cultural heritage "est très important, tant par sa richesse que par sa diversité et son ancienneté. Aujourd'hui, il est

[13] <http://www.irex.org/programs/MSI_EUR/2006/albania.asp> accessed March 22, 2008.

[14] <https://www.cia.gov/library/publications/the-world-factbook/geos/al.html> accessed March, 22, 2008.

vulnérable et menacé dans le contexte d'une société et d'un système économique en mutation." [15]

Since 1990, the writing style and language have also changed together with Albanian society. However, written texts are most often products of the intelligentsia, and therefore, we first need to define who is part of this elite in today's Albania. As Labov (2002) stresses, "it is the culturally dominant groups of society that are normally in the lead." The new urban intellectual elite is a generation that has lived through the previous regime, its fall, the transition period and that is acquiring today an ethical and political legitimacy. The individuals that are part of the elite have different backgrounds and personal stories ranging from members of the ruling caste to families fallen in disgrace through series of unfortunate events.

A distinctive feature of this class is the survivalist instinct. Many of them are well educated (often abroad) and they were able to camouflage their ideas as Ismail Kadare did in his voluminous literary oeuvre (Champseix, 2006). Some were not that successful and were persecuted or even imprisoned by the communist regime. There is also a new stratum created in the last decade which includes mostly people that have created substantial wealth and try to become part of intelligentsia by buying or creating their own media channels. This heterogeneous characteristic of the Albanian intelligentsia creates an inner conflict that can

[15] [Albanian cultural heritage] is very important for its richness, its diversity and its antiquity. Today, this heritage is vulnerable and threatened in the context of a society and economic system in transformation.

be seen, read and heard in all Albanian intellectual circles. A recurrent theme in these circles is the role intelligentsia played during transition with their cause célèbre, Ismail Kadare.

According to Coseriu (1997:31) language is a universal activity exercised individually while observing historical norms. In this definition can be identified a universal level, a historical level and the level of text (realized in either written or spoken forms). We will observe Albanian language realized in texts written by Kadare in two different realities. Even though Kadare's characteristic style has always allowed for an ambiguous reading and understanding (Champseix, 2006), his actual choice of words has changed according to the period in which he wrote. The two texts are *Koncert në fund të dimrit* 'The Concert at the End of Winter' (1988) and *Lulet e ftohta të marsit* 'Spring Flowers, Spring Frost [lit. The Cold Flowers of March]' (2000) representing writings before and after 1990 and both referring to a period towards the end of winter.

In *Lulet*, Kadare uses about 7,000 unique words out of a total of 35,000. In *Koncert*, he has used 17,000 unique words out of total of 174,000. Since the lengths are not the same, the counts for specific words are normalized to occurrences per one thousand words.

One of the first things noticed is the main product of the post-communist era, i.e. pluralism: while in general the number of singular words is more or less equivalent to the plurals, in *Koncert* the word *partia* 'party' in singular has a 50-to-1 ratio against its plural counterpart. In *Lulet* instead this ratio is only 3-to-2 corresponding perfectly to the ratio found in the 20 million words corpus of Albanian. But, most interesting is the list of collocations created by the word *partia* in Koncert. The top collocations include *komitet* 'committee,' *tanke* 'tanks,' *mbledhje* 'meeting,' *anëtar* 'member,' *politike* 'political,' *shtetit* 'of the state,' *puna* 'work,' and *ministri* 'the minister.' These words illustrate the role of the Party

during the communist era which dominated all aspects of governance from ministers to the military. In Lulet, the word does not appear enough to create any collocations.

The several variants of the word *punëtor* 'worker' have disappeared from 1988 to 2000 even though *punë* 'work' has remained exactly the same, at least from the statistical point of view. The worker or the proletarian was the symbol of the communist regime. It also appears that while *komunist* 'communist' or *vras* 'kill' and its variants have remained the same, *pushkatim* 'death penalty by shooting' has disappeared. In today's Albanian world, *komunist* is used to denigrate someone as often as it was used to glorify someone before 1990. The word *komunist* in Koncert cooccurs with *vend* 'country,' *është* 'is,' *njeri* (man.) In Lulet we find near *komunist* only *bërë* 'became,' *kishte* (had,) both past tense vs. the present tense in Koncert. People's smiles seem to matter more in *Koncert* which has twice as many variants of *buzëqesh* 'smile' while *Lulet* has many variants of *mort* 'funeral ceremony' that do not exist in *Koncert* even though both novels deal with deaths on more than one occasion. In addition to the higher number of occurrences, the word *buzëqesh* in Koncert co-occurs with several words such as *fytyrë* 'face,' *hidhërim* 'sadness,' *ironi* 'irony,' *përmbajtur* 'contained,' *fajtor* 'guilty,' and with the names of several characters (*Viola*,*Viktori*, *Silva*, *Linda*). In Lulet, it only appears with *mezi* 'hardly,' *vete* 'himself,' and the main character *Marku*.

Of course the words *internet* 'Internet,' *gomone* 'a very fast inflatable boat used for trafficking' and *Amnesty International* only appear in *Lulet* but together with them there are a lot of sexual references such as *seks*, *incest* and *homoseksual* that never appeared in *Koncert* (even though there are several passages that describe sexual encounters), as well as religious terms such as *Shën Mëria*, *Juda*, *Allah*, *Jehovah*, and the very common spoken expression *pash zotin* 'for God's sake' that was a taboo before. Political correctness was more prevalent during the

communist years: the pejorative word *jevg* 'Gypsy' finds its way in Kadare's prose after 1990 but not before that. Expression of racism appears to be a product of freedom.

Christian Leteux (2006) has compiled a significant list of neologisms introduced in the post-communist Albanian vocabulary which he groups in borrowings needed for new notions and objects, new meanings of existing terms or new compounds using already existing terms. He adds also a class of new constructs, mostly invented by the media for expressions borrowed from other languages.

Economy and politics constitute the longest lists with terms such as *tender*, *privatizoj*, *autostradë*, *kompetitivitet*, *elektorat*, *sondazh*, *pluralist*, *axhenda*, *lider*, etc. It is interesting to notice a higher number of terms borrowed from French which can be explained by the fact that most of the intelligentsia coming out of the communist years and that took control of the political process was educated in France. The rest are borrowed from English and Italian.

The more familiar terms that have to do with domestic objects take their names from the country of origin, usually Italy: *celular* (which in spoken language is being substituted by *cërre*), *pançetë*, *skadencë*, *lavapjatë*, *vetratë*, etc. The same can be observed in the show business which borrows heavily from the terminology used in Italian TV: *goleador*, *sfilatë*, *sipar*, etc. Some of the syntactic borrowings are expressions like *gërvisht dhe fito* (gratta e vinci), *është prononcuar* (s'est prononcé), etc.

The differences like the ones observed in Kadare's prose pre and post 1990 were supposed to be reflected in the revised Dictionary of Today's Albanian published in 2003. The main purpose of the new edition was the removal of any ideological influence. What is left in the Dictionary shows many of the internal conflicts of today's intelligentsia. While a

serious effort is made to purge the revised Dictionary of the communist ideology, fifty years of atheist propaganda still influence its editors. For example, in page 109 the entry for Bible still describes the Old Testament as a set of lessons and *myths* for the Hebrews and Christians. However, many of the neologisms identified by Leteux, are not yet included in the dictionary.

In Kadare's texts, as in the Dictionary, a few things have remained the same from the statistical point of view: between *kafe* and *kafene* and their related forms, in both texts there are about ten occurrences for every 1,000 words and they co-occur with words such as *zyrë* 'office,' *qendrës* 'of the center,' *hyrë* 'to enter,' *kthye* 'returned,' *Tiranës*, and *Pekin*. This is a statistically significant number. The same is true for *fjalë* and *flas* that co-occur with *përzemërt* 'cordial' in Koncert or *ëmbla* 'sweet' in Lulet and in both *shqiptoi* 'pronounced clearly' completing the typical Albanian *kafene* 'coffee house' ambiance with four occurrences for every 1,000 words. The *kafene* scene has not changed much in this decade and a half.

Another aspect of the attitude of intelligentsia during the communist regime was to stigmatize as provincials and unsophisticated the users of dialects especially in written materials. However, the truth is that there has been a large body of literature written in the Gheg dialect either from central Albania (Elbasan) or from northern Albania (Shkodër). The Shkodër dialect was also connected to the language of the Catholic Church and therefore was the butt of jokes or simply attacked by the communist propaganda for decades. Shkodër's intellectuals decided that they should declare Standard Albanian (based on Tosk) as a communist contraption. It was a rushed misstep that lead to a rift with Kosova's Gheg intellectuals and that therefore isolated the proponents from Shkodër. Since it is true that today's written Albanian has a serious disconnect with spoken Albanian, it would be much

more productive to study and offer solutions to this problem without bringing back ideology in language policies.

Besides the introduction of new words and the new pride in using dialects there are new themes introduced by authors young and not so young. What has made more news and sold more books has been the introduction of eroticism and the corresponding language. Fatos Kongoli, a former teacher of mathematics has been the leader. His books have been translated into several languages and get rave reviews in European literary magazines. The language used in this genre goes from outright vulgar to sophisticated word plays. This aspect of life too was a taboo for many decades.

As many taboos fell at the same time, as many new concepts are introduced, as is usual in times of revolution, language is reflecting these profound changes. In Labovian terms, we are faced with the phenomenon of *structural diffusion* between languages and dialects in contact caused by the sudden changes in Albanian reality. According to Mioni's macrocauses of language change taxonomy, Albanian reality is faced with two kinds of catastrophic events. On one hand, Albania is faced with mass migrations: from villages to the cities with depopulation in the mountainous areas and overpopulation in the areas around the capital, Tiranë[16] and from Albania to other countries[17] with a significant "brain

[16] In the period 1992-2002, the population in the region Tiranë-Krujë-Durrës has grown by 45%. Over the same period the population density has grown by 40% in the district of Tiranë and 43% in Durrës. Source: UNDP Albanian Human Development Report 2002 <http://hdr.undp.org/en/reports/nationalreports/europethecis/albania/albania_2002_en.pdf> accessed March 25, 2008.

drain." The sociopolitical and economic catastrophes include a new organization of government, the appearance of both nationalism and regionalism, and the return of religion. The compressed timeline of this process gives it the appearance of a sudden unshackling of language.

---

[17] By 2004 approximately 25 percent of the total population, or over 35 percent of the labor force, has emigrated. The country has approximately 900,000 emigrants, now residing mainly in Greece (600,000), Italy (200,000), and most of the remainder in other Western European countries, the US, and Canada. In the period 1990-2003, approximately 45 percent of the professors and researchers at universities and institutions emigrated, as did more than 65 percent of the scholars who received PhDs in the West in the period 1980-1990.
<http://www.migrationinformation.org/Profiles/display.cfm?ID=239> accessed March 25, 2008.

Deixis in Albanian

Albanian has an interesting set of similar pronouns and adverbs that form minimal pairs based on affix differentiation. The pairs are preceded by either *a-* or *k(ë)-*. While Xhuvani and Çabej (1976) treat *a-* and *k(ë)-* as prefixes, in other studies, Çabej (1976a) refers to them as deictic particles. In contemporary literature, *a-* and *k(ë)-* are considered morphologically overt semantic components which are called features (Fillmore, 1982 and Diessel, 1999) or parameters (Pederson and Wilkins , 1996). In the paradigm of Albanian deictics, we include primarily demonstrative determinatives, demonstrative pronouns, and deictic locative adverbs. Since the pair *a-* and *k(ë)-* also appears in some other demonstrative adverbs, these adverbs are also included in this study.

The difference in treating *a-* and *k(ë)-* as prefixes or as particles is important when trying to date the creation of this dichotomy. Historical phonetic patterns in Albanian would have caused the deletion or reduction of the prefix *a-* to *ë-* /ə/. The fact that it has not happened would mean that these could be new constructs. If instead we were dealing with a particle, the chances would be that an originally independent word would not be reduced or deleted. Of this opinion is Vladimir Orel in his analysis of all the entries for these words in his etymological dictionary (Orel, 1998). Eric Hamp (p.c.) also suggests that *a* was a separate word and is etymologically related to the Old Irish privative *ao* reflected as *a* in Old Albanian.

The other component of this dichotomy, *k(ë)-* is more transparently related to the common Indo European proximal particle **ko* that Bühler (1934/1990:126) calls the **ko-* stem of the here-deixis. The *satem* classification of Albanian would dispute this lineage but, as

Friedman (2003; based on Hamp) illustrates with the three Proto Indo European velar reflections below, **k* can remain *k* in today's Albanian[18]:

```
*kʹē-t-    kʹ > th / __ ē   thotë 'says'
*kēska     k > k / __ ē     kohë 'time'
*kʷērsna   kʷ > s / __ ē    sorrë 'crow'
```

Another factor that supports the historical derivation from two different words is the existence of the pronominal roots *të*, *ta*, *to*, *tij*, *saj*, *tyre*, *tillë*, *tilla* and the substitution of the particles *a-* and *k(ë)-* with *nji-*, with an article or with a preposition as in *njita* 'those,' *me to* 'with them,' *e tyre* 'of them,' etc. Theoretically, this derivation follows the cline "phrases or words > non-bound grams > bound grams" (Brinton and Traugott, 2005). Frequency and repetition of these forms close to each other brought about their eventual merging into one (Bybee and Hopper, 2001).

The table below contains an extensive list of these words with their respective number of occurrences in the Corpus of Albanian Language Text (CALT from now on). Even though some of them have very low or zero frequencies in CALT, they have been verified for usage examples in grammars, dictionaries, various texts and in the Google search engine.

category	a+___	freq	k(ë)+___	freq
adverb of direction	andej	901	këndej	233
adverb of quantity	aq	7081	kaq	4140
adverb of manner	ashtu	9125	kështu	8167
adverb of place	atje	4615	këtje	1
adverb of place	aty	5855	këtu	6789
demonstrative singular masculine NOM	ai	43310	ky	17502

[18] Paolo Di Giovine (p.c.) provides a more detailed rule set where *k (velar) > k; *k (palatal) > th; $*kw_{a/o} > k_{a/o}$; $kw_{e/i} > s_{e/i}$.

demonstrative singular feminine NOM	ajo	19890	kjo	26399
demonstrative singular ACC	atë	23200	këtë	62972
demonstrative plural masculine NOM/ACC	ata	22803	këta	4610
demonstrative plural feminine NOM/ACC	ato	15865	këto	21528
demonstrative singular masculine DAT/GEN/ABL	atij	4219	këtij	22809
demonstrative singular feminine DAT/GEN/ABL	asaj	4604	kësaj	20048
demonstrative plural DAT/GEN/ABL	atyre	5319	këtyre	9686
demonstrative pronoun plural masculine ABL	asish	0	kësish	0
demonstrative pronoun plural feminine ABL	asosh	1	kësosh	0
demonstrative determinative masculine ABL	asi	32	kësi	80
demonstrative determinative feminine ABL	aso	47	këso	68
i/e/të/së + demonstrative adjective 'such'	atillë	85	këtillë	119
i/e/të/së + demonstrative adjective plural feminine	atilla	35	këtilla	38

Table 5: Words containing one of the deictic affixes

The deictic words in Table 5 are divided in four groups representing adverbs, words that can be both pronouns and determinatives (as in *ky është* 'this is' or *ky libër* 'this book'), pronouns only (as in *disa asosh* 'some of them'), and determinatives only (as in *kësi djemsh* 'this kind of guys'). All of them are stressed in the second syllable. As can be observed from the table the words that can only be pronouns are the masculine and feminine ablative forms of the demonstrative *asish/kësish*, *asosh/kësosh*. They do have their corresponding determinative form *asi/kësi*, *aso/këso*. Because of the functional similarity (Likaj, 1984) in their syntactic roles, a noun declension device (the *–sh* ending) which includes ±definiteness has been extended by analogy to the pronoun. The very low number (sometimes zero) of occurrences corresponds with the continuous pressure from the use of the prepositional form and the odd creation of the category of ±definiteness in Albanian pronouns. Just as usage frequency which became known as analogy (Smith, 2001) must have created these forms, the increasing frequency of the sequence *PREPOSITION + atyre* is making them disappear.

Albanian includes also a peculiar combination of the derivational devices mentioned above. As Çabej notes in his *Studime etimologjike në fushë të Shqipes* (1986:245), through

morpho-etymological analysis, the word *tutje* is found to contain both primitive roots for 'here' (*tu*) and for 'there' (*tje*). Other words contain both distal and proximal deictic particles that create the indefinite meaning of neither close nor far. Even though they never occurred in CALT, they are included in Table 6 to complete the view of this set of deictic words.

category		CALT occurrences
adverb of place	tutje	699
adverb of place	akëku	0
indefinite pronoun singular .masculine	akëcili	0
indefinite pronoun singular feminine	akëcila	0
indefinite pronoun singular	akëkush	0

Table 6: Words formed by both distal and proximal parts

The point of view that the prefix *akë-* could be a combination of both deictic particles is offered as one of the possible derivations by Çabej (1976a:32) and Demiraj (2002:259).

Discussion of the Method

Albanian language, which preserves some archaic features of the Indo-European languages, has a long history of etymological and grammatical studies, but the new capabilities offered by today's powerful computers have not yet exploited this history. This study establishes the application of computational techniques to Albanian. Roberto Busa, a pioneer in linguistic text analysis, often says that the computer allows and, at the same time, requires a new way of studying languages. Using "state of the art" computers, he started since 1949 his efforts to search out new meaning in verbal records, to view a writer's work in its totality, to establish a firmer base in reality for the ascent to universal truth (Raben 1987). The approach that he calls "*nuova* filologia" and that is based on the analysis of extended universes of text, allows for a new start that would conduce to the redefinition of grammatical and lexical categories, thus improving grammars and vocabularies (Busa 1987). To describe a language, even just its grammar, one can only start from the text (Coseriu, 1997:56). While contextual analysis or what Coseriu calls "transphrastic grammar" deals with text organization and corresponding operations in small amounts of texts, our approach's intent is to find similar relations through massive text analysis. Indeed, it is common now to approach theoretical issues with support of large amounts of naturally occurring text. As Joan Bybee (2006) notes, this practice has been in place for decades in the works of those who examine how discourse shapes grammar from both synchronic and diachronic points of view. The most natural approach for this kind of analysis is to look for patterns created through repeated usage of linguistic tokens. She lists three outcomes deriving from frequency of usage:

i. low levels of repetition lead to conventionalization only (as in prefabs and idioms)

ii. higher levels of repetition can lead to the establishment of a new construction with its own categories

iii. extreme high frequency leads to the grammaticalization of the new construction, the creation of grammatical morphemes, and changes in constituency

Frequency of collocations is seen as affecting both morphosyntactic and semantic representations. From the cognitive point of view, similar tokens appearing in similar constructs start creating patterns that go beyond the simple occurrence of sequences of some particular strings. Such patterns lead to the search for more abstract repetitions that are observed beyond the surface representation. This study aims at discerning relationships between morphosyntactic and semantic categories or tokens through their usage in large amounts of texts and how these relationships in turn affect the very grammar that governs their usage. Interdependence among words not only determines the choice of words – it also determines how words are produced: for example, lexical frequency and word-to-word probability suggest that probabilities are capable of affecting pronunciation. Studying /t,d/-reduction in American English, Jurafsky et al. (2001) have shown that the probability that a word appears after a certain word $P(w_i|w_{i-1})$ affects the amount of reduction in proportion to its predictability and that this reduction is not necessarily due to lexicalization. Gahl and Garnsey (2004) go beyond word-to-word transition: through several arguments and experiments they conclude that grammar should include probabilities of syntactic structures as well. Accounting for such probabilities, such as verb biases for direct object or sentential

component, creates knowledge that affects language production and comprehension thus forming an integral part of grammatical knowledge.

Based on Firth's (1957a) conclusions that "meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words" as in the example where "one of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, its collocation with *night*," this study follows his slogan discovering new knowledge about words from the company they keep (Firth, 1957b). As collocations constitute a central role in this approach to lexical-grammatical knowledge discovery, what constitutes a collocation has to be defined clearly. Christopher Manning and Hinrich Schütze in their Foundations of Statistical Natural Language Processing (1999) provide a list of criteria that define collocations, i.e. non-compositionality, non-substitutability and non-modifiability which are closer to Choueka's (1988) definition as a "syntactic and semantic unit whose exact and unambiguous meaning cannot be derived directly from the meaning or connotation of its components." But that is a different kind of collocation that helps in discovering tight semantic pairs of usually content words that range from prefabs or idioms to grammaticalized sequences. According to Firth (1957b), the collocation of a word is not to be regarded as mere juxtaposition, it is an order of *mutual expectancy* or more specifically, the elements of the meaning of the word are indicated when their habitual word accompaniments are shown (Firth 1968). In the same vein, in his numerous publications on Aktionsart, Pier Marco Bertinetto (1986, 2001 and 2007) considers the compatibility analysis between verbal classes and selected sets of adverbials as a very efficient tool for outlining syntactic patterns in the usage of tenses. At the same time that grammar influences usage, usage also influences grammar. Joan Bybee's (2006) conclusion that "just as there can be no discrete separation of grammar and lexicon because

there are so many cases in which specific lexical items go with and/or require certain grammatical structures, so also there can be no strict separation of grammar and usage" coincides with the conclusions we reach by studying the relationships in which Albanian deictics enter. While pronouns and adverbs are usually considered function words, the presence of deictic features adds to them a meaning that might affect surrounding words or be affected by them and even by extra-linguistic information. The focus therefore remains on the constellation of the strongly associated words surrounding the target and how they define each other in the way Firth's *dark* and *night* do. Just as agreement (or what Firth calls colligation) determines the syntagmatic relationships between words, there are elements that determine the ability of a word to be collocated only with certain other words. We look for similar abstract elements that allow words to enter into collocational relationships. These elements can be not only semantic but also morphosyntactic such as case, number, tense, and person. In our approach we look for elements (semantic, grammatical or something else) that affect the usage of words containing deictic elements and how usage, therefore, defines the meaning of these deictic words.

This analysis is performed on a newly built nineteen million word corpus of texts from today's Albanian. Collocations are extracted from pairs of words occurring within a given window of words in the text. A pair is made up of a target word and a collocated word. The target words in our case are all words that contain a deictic particle. Every pair is tested whether the presence in each-other's neighborhood is random or there is some gravitational force that creates the constellation. The text window size is determined by a left and right threshold distance inside which the words are inspected. The extracted collocations are then tagged with semantic and/or morphosyntactic information and analyzed for relevancy.

Corpus building

Before starting any collocational analysis, the first step is assembling a suitable corpus and tools for exploring it. Quantitative corpus based analysis of Albanian is still in its initial phases. The efforts towards creating a balanced corpus have been unsuccessful and there are no accessible corpora for the research community. Another issue with the Albanian language is the relatively young age of the standardized language. The two main dialects, Tosk and Gheg, remain very much in use, confining the standard mostly to the written language. After the fall of communism in the early 1990's, new concepts, both technical and social, were introduced. The language has reacted with the introduction of newly created terms from internal resources or direct foreign loan words. So, standard Albanian is now in a significant state of flux and under pressure from the dialects on one side and aggressive borrowing on the other.

Deictic pronouns and adverbs, which are the object of this study, are functional words. As such, their usage is highly grammaticalized and therefore less likely to conscious variation. Consequently, the quality of collocations for such words should not be affected by temporary linguistic trends. However, the corpus needs to represent today's language in its entirety and not become simply a collection of texts (Biber et al. 1998). With Coseriu (1997:49), we see text as structured in several layers consisting of phrases, clauses, syntagms, words and meaningful elements. We are interested in the relationships the smaller elements enter when organized in one of the higher layers. Even though the intended investigation has a very narrow scope and mostly deals with function words, every effort has been made to make the corpus as representative as possible. The study of collocations requires massive amounts of texts because of some low frequency phenomena or phenomena appearing only

in some text types. For this reason, we expanded the corpus well beyond the usual few million words.

The corpus used for this study was created by extracting content from several Internet sites and scanned material. The sites were selected following criteria of quality and content. The text contained in these sites had to be written in standard Albanian following Albanian orthographic rules and using the correct characters. These criteria eliminated most of the Albanian language Internet lists where Albanian is mixed with other languages and where writers almost never use the diacritics marks for *ë* and *ç*. As for the content, an effort was made to balance news items with literary prose and interviews. In addition to online newspapers, several sites on literature and culture were included in the spider list and they were regularly mined between 2003 and 2005. Church and Mercer (1993) prove in the introduction to the Computational Linguistics' Special Issue on Using Large Corpora that "more data is better data" and conclude that "it is not required that the corpora be balanced, but rather that their quirks be uncorrelated across a range of different corpora." In our corpus building we abide to "more data is better data" principle. To achieve growth of the size of the corpus, some literary works from well known authors as well as some historical and philosophical books and articles scanned or already in electronic form were included. Indeed, as the corpus grew, the quality of collocations improved as well.

Content acquired from the Internet sites required careful handling. Every downloaded page has been analyzed and cleaned by a page scraper, removing HTML tags and template elements. Obviously, the template text, repeated in every page from the same site, would distort the counts and diminish the statistical accuracy. The most salient example is the word *këtë* 'this' which has a count of 215,000 in Google (accessed in March 2005).

However, 19%, or 40,500 instances, are part of the phrase *këtë faqe* 'this page' or some other constructs like it that point to the page that contains it. These kinds of phrases usually appear in the template elements and eliminating them would prove beneficial to our collocational analysis. The remaining content after the clean-up is saved as text only and indexed for quick searching.

Having the data indexed provides a simple tool for eliminating duplicates. A sequential string of ten to fifteen words from every new page is submitted as a query to the search engine. If there is a 100% match, the new document is considered a duplicate and not stored. Obviously, there is the risk of eliminating texts that quote each other but in our data the quantity of eliminated text did not constitute a problem. The collection consists now of approximately 19 million tokens and 250,000 types.

Computational tools

The tools for analyzing the corpus include a tokenizer, indexer, concordancer, collocator, set computation utilities, and a search engine allowing the use of regular expressions. All these utilities are written in Java.

The tokenizer is configurable and uses rules specific to Albanian. There are also specific rules for sorting the collocations following the 36-letter Albanian alphabet where:

```
a > b > c > ç > d > dh > e > ë > f > g > gj > h >
i > j > k > l > ll > m > n > nj > o > p > q > r >
rr > s > sh > t > th > v > u > x > xh > y > z > zh
```

All non-letter characters such as punctuation, blank or other symbols are set as word boundaries.

Mathematical apparatus

Collocations, as defined above, are computed by using Mutual Information (MI) as defined by Church and Hanks (1991) as well as Z-score and T-score as defined by Barnbrook (1996) and implemented in Mason (2000). These computationally different scores measure the amount of attraction words exert on other words and rank the pairs for further analysis. The first step is to define the object of these computations. The procedure consists of a first step that includes the selection of all the concordances of the nucleus word also known as target word. The object words are defined by the window size: four words to the left and four words to the right of the nucleus word. According to Sinclair (1991), beyond four words from the nucleus, there are no statistical indications of the attractive power of the nucleus word. These make up the neighborhood of the nucleus word or, all together, a nucleus word centered sub-corpus.

word	…	word	**word**	**word**	**word**	**word**	*nucleus*	**word**	**word**	**word**	**word**	word	…	word
-N	…	-5	**-4**	**-3**	**-2**	**-1**	*0*	**1**	**2**	**3**	**4**	5	…	N

Figure 5: Concordance lines with highlighted nucleus and object words

The MI-score is the ratio of the probability that two given words appear in each other's neighborhood with the product of the probabilities of each of the words in the corpus. It could also be considered as the ratio of what is observed with what is expected. Expectation is also called the hypothesis zero: two words can appear in each other's neighborhood by pure chance.

$$MI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

$$P(w_1, w_2) = \frac{count(w_1, w_2)}{N}$$

$$P(w_1) = \frac{count(w_1)}{N}$$

$$P(w_2) = \frac{count(w_2)}{N}$$

$$MI(w_1, w_2) = \log_2 \frac{\frac{count(w_1, w_2)}{N}}{\frac{count(w_1)}{N} * \frac{count(w_2)}{N}} = \log_2 \frac{count(w_1, w_2) * N}{count(w_1) * count(w_2)}$$

$$N = Size\ of\ Corpus$$

Equation 1: Mutual Information computation

The MI-score indicates how much more certain we are that when we find one of the words we will also find the other one. Certainty is expressed in the number of bits that increase the information of each of the words when the other is found (Manning and Schütze, 1999). However, we notice that, from the formula above, the mutual information increases with the rarity of the event and it gravitates around zero when two words are completely independent. Church and Hanks (1991) propose an arbitrary threshold for rare collocations which is set to five in our computations. Everything that appears less than five times is ignored but this does not always help. For example, the word *punceura* has a total of six occurrences in our corpus. But all these six occurrences appear close to *këtu* 'here' in the reported letter of the death of Scanderbeg *u sëmur këtu nga punceura* 'got sick here from the *punceura*' creating a pair that occurs just enough times to be above the frequency threshold.

This makes *punceura*, a word[19] that is not found in any Albanian dictionaries, as the top collocated word for *këtu* with the probability of appearing together 2,800 times greater than chance (MI=11.45 - > $2^{11.45}$=2800). While rarity of the event influences heavily the dependence between words, it does not do so for independence. It can be concluded that mutual information alone is a good measure of independence but not very reliable for conveying dependence.

Another measurement, the T-score, confirms that the high MI-score is not created by just two rare words that happen to appear close to each other. Church et al. (1991) conclude that MI is better for highlighting similarity and T-scores are better for establishing differences.

$$T\_score = \frac{observed - expected}{\sqrt{observed}}$$
$$observed = count(w_1)$$
$$expected = p(w_1) * n$$
$$p(w_1) = \frac{count(w_1)}{N}$$
$$N = Size\ of\ Corpus$$
$$n = Count(Nucleus\ Word) * Window\ Size$$

Equation 2: T-score computation

The "expected" value predicts the count of collocations inside the sub-corpus that the nucleus word creates through its concordance lines. The "observed" value is the actual

[19] This could be a typo in the translation or in the scanning of the translation. A common error made by scanners is to misread an italic font *t* as *e* [*punctura* > *punceura*] which could mean that Scanderbeg died from malaria caused by a punctura (mosquito bite) and not poison as some say.

count. A T-test could be used to compare the hypothesis that two words appear near each-other too often for the sequence to be a fluke against the null hypothesis that the observations can be attributed to chance. The T-score result can be interpreted as a number of standard deviations. Theoretically, if the T-score is larger than 1.65 standard deviations, then we ought to believe that the co-occurrences are significant and we can reject the null hypothesis with 95% confidence (Church & Mercer 1993). Rankings based on T-score are different from MI but, looking at listings it seems that if both computations generate a high ranking for a certain word, then that word appears intuitively to be a more appropriate collocation.

The top twenty collocations of the word *këtu* 'here' are lead by the obscure word *punceura* followed by the typo *përshirë* (for *përfshirë* 'included').

collocate	MI	t-score
punceura	11.450145	2.442485
shto	10.587648	3.299382
përshirë	9.642790	2.424976
ndodhem	9.227753	3.417879
përfshi	9.214631	9.148748
sillni	9.096508	2.956158
rrihet	8.935572	2.602523
erdha	8.897173	6.060979
erdhëm	8.575676	3.391466
eja	8.543254	3.661419
përjashto	8.511545	2.395794
shtojmë	8.450145	2.585232
rrimë	8.320862	3.377435
mbaron	8.314368	6.009494
vërejtjen	8.170037	2.572264
përfshirë	8.110012	17.234327
ndalen	8.087575	2.567941
nejse	8.087575	2.567941
shtoj	8.050214	2.909457

Table 7: Collocations for *këtu* ranked by MI-score

If the same data is sorted according to the T-score, a new view is obtained that appears more acceptable to the linguistic intuition.

collocate	MI	t-score
përfshirë	8.110012	17.234327
jam	6.299317	12.318520
edhe	3.969683	12.096745
kam	5.224640	11.525725
unë	4.973386	11.474445
jemi	6.108536	11.252233
ne	4.507155	11.151979
pikërisht	6.059004	10.633191
nuk	3.720507	10.467521
por	3.965955	10.387787
ardhur	5.855198	10.123889
që	3.477561	9.758098
shqipëri	4.909296	9.164880
përfshi	9.214631	9.148748
deri	4.496521	8.733419
kemi	4.840214	8.678616
aty	5.419602	8.566459
po	3.943463	8.202535
ju	4.792727	7.985154

Table 8: Collocations of *këtu* ranked by T-score

Another way for measuring collocation strength uses standard deviation. Known as Z-score, it makes a comparison in relation to the mean and standard deviation. This score determines the relative importance of specific items within given collections.

$$Z\_score = \frac{observed - expected}{std}$$

$$std = \sqrt{n * p(w_1) * (1 - p(w_1))} = \sqrt{n * \frac{count(w_1)}{N} * (1 - \frac{count(w_1)}{N})}$$

$$N = Size\ of\ Corpus$$

$$n = Count(Nucleus\ Word) * Window\ Size$$

Equation 3: Z-score computation

This formula is interpreted more easily: a Z-score is the ratio between an observed deviation and a standard deviation (Muller, 1973.) Rankings based on Z-score are more

similar to T-score but give a different view of closed class words since its computation is based on standard deviation and the closed class words tend to have constant distribution.

collocate	MI	t-score	z-score
përfshirë	8.110012	17.234327	101.289470
përfshi	9.214631	9.148748	78.842710
erdha	8.897173	6.060979	46.790560
shto	10.587648	3.299382	45.760193
punceura	11.450145	2.442485	45.677853
jam	6.299317	12.318520	38.657307
mbaron	8.314368	6.009494	37.908257
jemi	6.108536	11.252233	33.051810
pikërisht	6.059004	10.633191	30.701508
ndodhem	9.227753	3.417879	29.588812
vij	8.021301	4.942011	28.163593
ardhur	5.855198	10.123889	27.237911
kujtojmë	7.865182	4.924066	26.583393
eja	8.543254	3.661419	25.003086
kam	5.224640	11.525725	24.928375
sillni	9.096508	2.956158	24.453670
përshirë	9.642790	2.4249766	24.240822
përmendur	6.923668	6.1253304	23.862272
erdhëm	8.575676	3.3914669	23.421322

Table 9: Collocations of *këtu* ranked by Z-score

By observing Table 7, Table 8, and Table 9, it can be concluded that a combination of all three measures would provide the best linguistic quality rankings of the collocations. Since, mathematically, they cannot be combined, we follow a more pragmatic solution which consists of averaging the ranks obtained using each score and sorting based on the average ranking value. This sorting floats at the top of the list words that define the main meanings and usage of the analyzed word.

collocate	mi-rank	t-rank	z-rank	combined
përfshirë	16	1	1	6.00
përfshi	5	14	2	7.00
erdha	8	33	3	14.67
mbaron	14	34	7	18.33
jam	66	2	6	24.67
vij	20	49	11	26.67
kujtojmë	22	50	13	28.33

jemi	74	6	8	29.33
përmendur	42	31	18	30.33
pikërisht	78	8	9	31.67
përmend	35	40	23	32.67
llogaritur	25	56	20	33.67
bëjmë	63	24	21	36.00
ardhur	88	11	12	37.00
eja	10	92	14	38.67
ndodhem	4	105	10	39.67
shto	2	117	4	41.00
përmendet	31	67	26	41.33
rri	41	52	31	41.33

Table 10: Collocations of *këtu* ranked by average rank

From a linguistic point of view, Table 10 contains the most intuitive data. Following this approach, rare words, boosted by the MI-score, and very frequent words, boosted by the T-score, are both ranked lower.

On the other hand, the bottom of the list is interesting as well: it contains collocations which infer inherent differences thanks to the negative values of T- and Z-scores.

collocate	MI-score	T-score	Z-score	MI-rank	T-rank	Z-rank	average
te	1.4807768	-11.655456	-6.891453	1013	1024	1023	1020.000
policia	0.8368158	-9.840164	-4.651082	1027	1020	1014	1020.333
brenda	0.5422527	-11.006984	-4.697560	1028	1022	1015	1021.667
2003	0.4777633	-11.622144	-4.850535	1029	1023	1016	1022.667
dje	0.9977749	-12.024696	-6.011262	1026	1025	1020	1023.667

Table 11: Bottom ranked collocations for *këtu*

In Table 11, the Mutual Information score for the words *këtu* and *dje* is close to zero. It means that *dje* 'yesterday' could as well be there by pure chance. But T- and Z-scores have large negative values (-12 and -6) which indicate that *këtu* influences negatively the number of occurrences of *dje*.

The argument of using the average rank for capturing the most significant collocations is illustrated in the four graphs below. In Figure 6 and Figure 7 have obvious

rough lines that show how much MI and T scoring collocations differ one from the other. Figure 8 and Figure 9 instead have much more smoother lines showing a certain concordance among the respective computations. While ranking using the Z-score appears to have a very smooth line, we found that some of the higher collocations using the MI-score are moved to the bottom of the list. By analyzing the effect of an average rank sorting, we found that most of the higher values by all three methods of computation remain at the top and the lower values are still at the bottom.

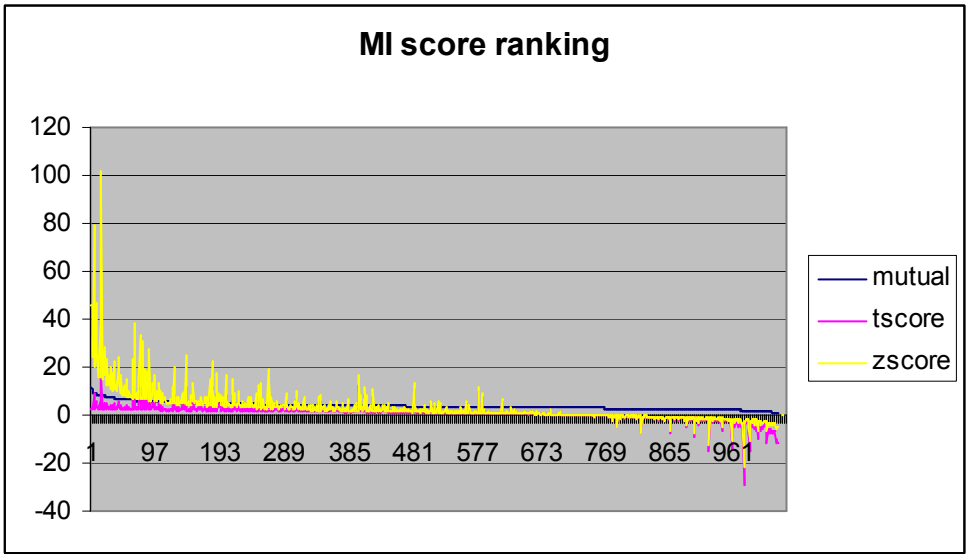


Figure 6: Graph of collocation scores for *këtu* sorted using MI values

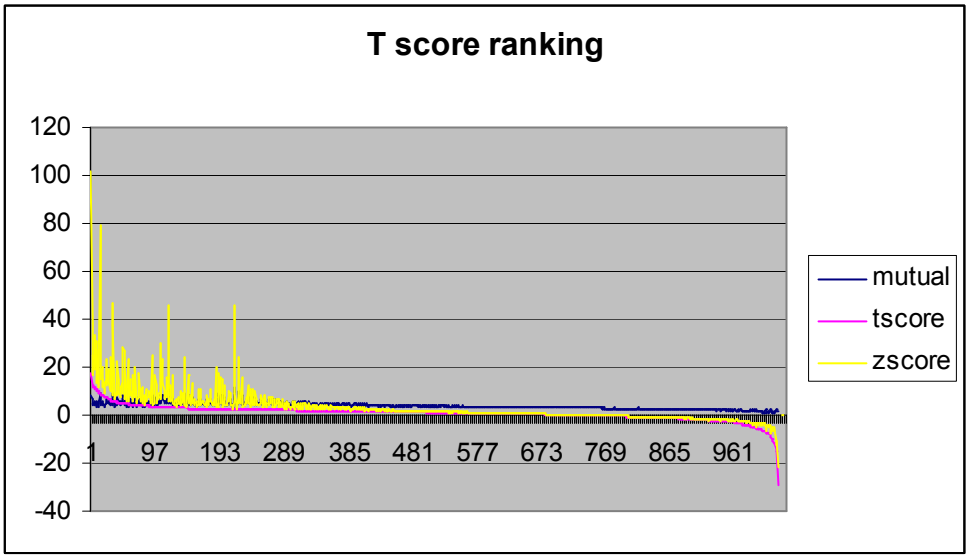


Figure 7: Graph of collocation scores for *këtu* sorted using T values

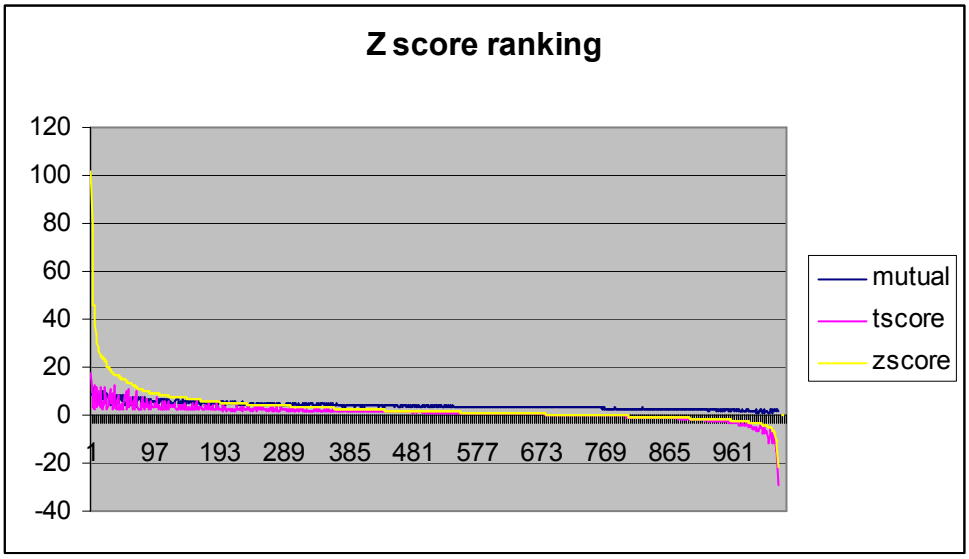


Figure 8: Graph of collocation scores for *këtu* sorted using Z values

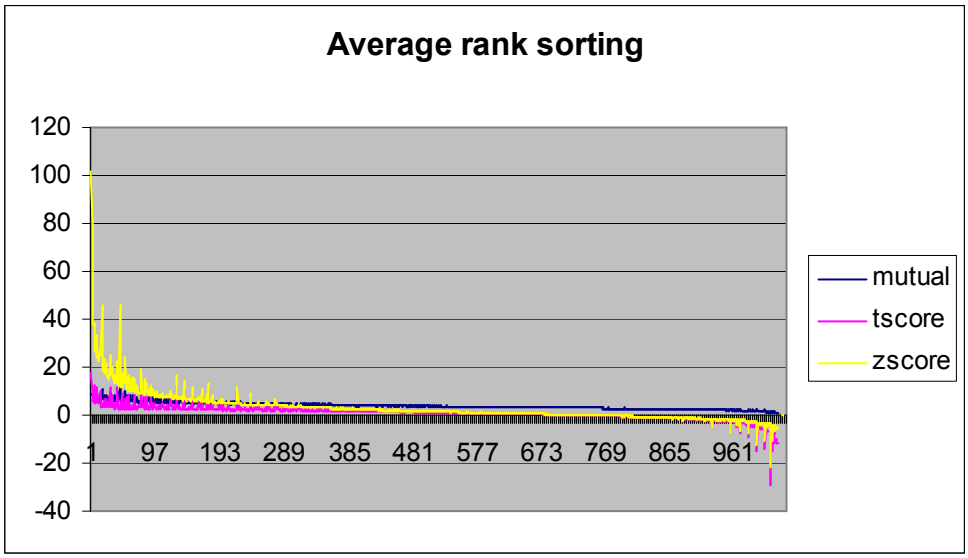


Figure 9: Graph of collocation scores for *këtu* sorted using average rank values

Another concept introduced by Sinclair (1991) is the upward and downward collocation. Upward collocates are those whose own occurrence is over 115 percent of the node and downward, less than 80 percent. The first group consists of words which habitually occur in language more frequently than they do themselves, e.g. *shkon* 'goes,' which occurs 1706 times in CALT, collocates with *në* 'in/at,' 656731 occurrences, *atje* 'there,' 5296 occurrences, both of which are more frequent words than *shkon*. Similarly, the 'downward' collocations are words which habitually occur less frequently than they do, e.g. words *pushime* 'vacations,' 378 occurrences, *shkollë* 'school,' 1361 occurrences, are downward collocations of *shkon*. Sinclair makes a sharp distinction between these two categories claiming that the elements of the 'upward' collocation (mostly prepositions, adverbs, conjunctions, pronouns) tend to form grammatical frames while the elements of the 'downward' collocation (mostly nouns and verbs) by contrast give a semantic analysis of a word. In our approach we make equal use of both kinds and look for significance in both directions.

New Philology

The wordlists resulting from this kind of analysis allow for abundant interpretation. Through rankings computed using the approach discussed above, word forms are grouped together allowing for further investigation: from comparison with dictionary definitions, to considerations from the grammatical point of view, word-formation, descriptive etymology, and phrasal stylistics (Firth 1957b). This approach, formalized as the idiom principle by Sinclair (1991) looks for the factors that cause the simultaneous choice of two words. With our analysis we can generalize this choice as the attraction caused by the respective bundles of grammatical, semantic and pragmatic features of each word.

Evidence from the collocational and statistical analysis of large amounts of texts shows that usage provides an improved and more reliable ranking of meanings from what introspective approaches would typically offer. For example, as Stefan Grier notes in his analysis (Grier, 2006) of the verb *to run*, by combining a cognitive/introspective approach with corpus data and syntactic constructs, the senses are clearly clustered around 'fast pedestrian motion' and 'to manage.' As texts, in general, are made up of occurrences of frequent words as well as frequent senses of the less frequent words, we will focus our analysis at the top ranked collocations.

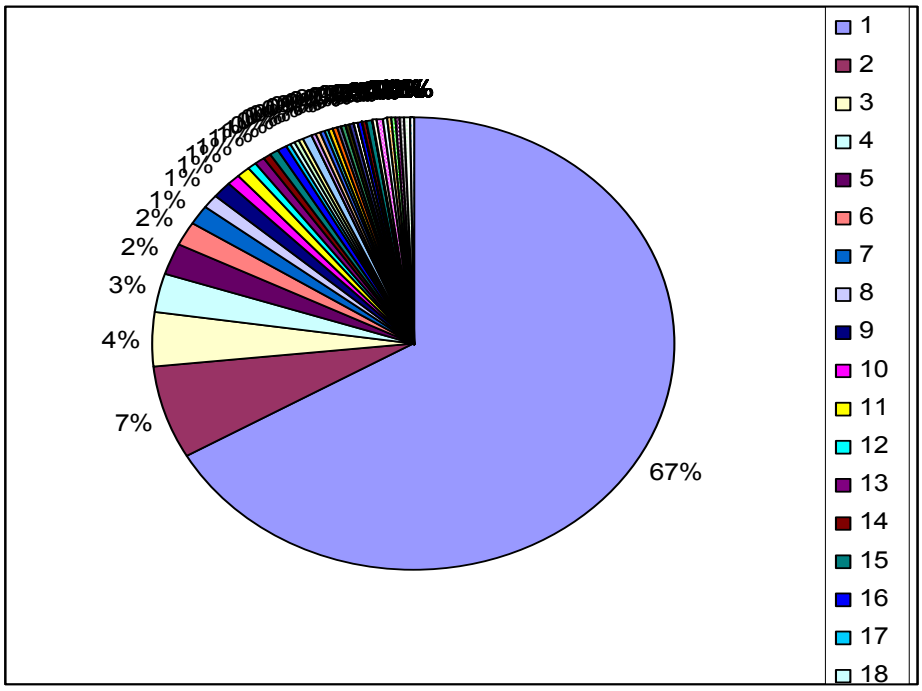


Figure 10: Distribution of tokens among types: first 1000=67%; second 1000=7%; third 1000 = 4%;...

Figure 10 illustrates the distribution of single words in CALT. The one thousand most frequent words make up more than two thirds (67percent) of the 19 million words (or tokens) produced by the 264,000 word forms (or types). Combining the different collocation rankings, we get a good combination of collocations made by frequent words as well as by frequent senses of rarer words. Starting with the observation that the most frequent patterns in language cover the majority of produced text, Abir et al. (2002) and Carbonell at al. (2006) have built a theory based on the repetitive and overlapping nature of structures of texts. By isolating chunks of text that are repeated often and are verified by valid overlaps with other chunks of texts, they have achieved very good practical results in Machine Translation and Text Mining. Beside the obvious Zipfian distribution of single words, they find that the same laws apply to strings made by multiple words. Looking for rather large chunks of texts or *n-grams* that can be found in repetitive patterns they extract units (just as Sinclair's *idioms*)

which can be further processed for translation, indexing or text refinement. Our analysis is based not only on sequences but also on distant words that have the same attraction towards each-other. We try to identify the features that make up each bundle and try to build a lexical-grammatical model of their cohesiveness. According to Sinclair (1991), a model of language which divides grammar and lexis, and which uses the grammar to provide a string of lexical choice points, is a secondary model… the other principle, the idiom principle, dominates.

Deictic Adverbs

Adverbs that are formed by the opposition of a- and k(ë)- in Albanian are the adverbs of place or locatives *atje/këtje* and *aty/këtu* 'there/yonder/here,' the directional adverbs *andej/këndej* 'thither/hither,' the adverbs of quantity *aq/kaq* 'that much/this much' and the adverbs of manner *ashtu/kështu* 'that way/this way.'

Among Balkan languages, we find similar and even more regular oppositions in Serbo-Croatian: *му/онде/овде*, *тамо/онамо/овамо*, *толико/онолико/оволико*, *тако/онако/овако*. Greek renders such concepts analytically *εδώ/εκεί*, *δώθε, προς τα εδώ/προς τα εκεί* and Romance languages either use analytical devices or do not have an opposition as in Romanian *aici/acolo* 'here/there,' *de aici/de acolo* 'hinc/inde,' *pe aici/pe acolo* 'hac/istac' but *aşa* 'this way,' *atât* 'this much,' etc.

Deictic adverbs of place

Standard Albanian has three deictic adverbs of place which are *këtu* 'here,' *aty* 'there,' and *atje* 'there.' All three forms are compounds consisting of a deictic prefix and a historically pronominal form. The deictic proximal prefix that forms *këtu* is *kë-* and the distal forming both *aty* and *atje* is *a-*.

The following table contains the number of occurrences of these words in CALT, their respective ranks, the numbers of collocations with words that occur more than five times, and two verification columns with the number of occurrences from Google and AltaVista search engines accessed on July 23, 2006. As of this date, AltaVista identifies approximately 1.2 million pages in Albanian language. Google's corpus is estimated to have about 6 million pages in Albanian. The numbers in the table mean that there are that many pages containing the search term and not the real number of occurrences.

word	occurrences	CALT rank	collocations	Google	AltaVista
këtu	6789	178	1029	564000	385000
aty	5855	198	849	248000	98800
atje	4615	279	758	118000	79700

Table 12: Distribution of deictic locative adverbs

Out of the 250,000 word forms that make up CALT, all three deictic locative adverbs are ranked quite high. Nevertheless, we observe that *këtu* has a higher frequency than both other forms, a fact that is reinforced by the much larger Google and AltaVista corpora where the ratios are approximately 3-to-1 for *këtu*/*aty* and 5-to-1 for *këtu*/*atje*. The difference in frequencies is one of the parameters that correlate with the concept of markedness or the degree of generality. This concept has been very effectively applied to phonological studies where it has been noticed that unmarked phonemes occur much more frequently than marked ones. Later, this concept was extended to grammar and lexicology.

Based on the markedness theory, the data for the Albanian deictic adverbs of place would conduce to the conclusion that the level of markedness is higher for the rarest form *atje* and that *këtu*, the most frequent one, is the least marked. This corresponds with the analysis by Benor and Levy (2006) where proximals are the least marked in the deictic adverbs series. But, from an opposing point of view, Mayerthaler (1988) in his proposed dichotomy between less and more marked semantic properties argues that *here* is more marked than *there* because one sees others more than oneself. Therefore, even in Albanian, *këtu* should be more marked than *atje*. While the logical analysis alone is inconclusive, further statistical observations add to this inconclusiveness. Henry Kučera notes that the markedness/frequency correlation does not hold in lexical and grammatical levels (1982). As stressed by Victor Friedman statistical rarity is not a defining characteristic of markedness. "It can and does happen in language that the meaning for which a word is marked is one which speakers have frequent cause to specify" (1994). Furthermore, Martin Haspelmath (2006) considers markedness superfluous and notes that the term only adds confusion. For all these reasons, it is impossible to conclude that frequency defines markedness in the case of these three adverbs. We will maintain frequency as a sufficiently discriminating feature of the deictics. There is also another point of view regarding the order of acquisition of deictic terms. Experiments (Tanz, 1980:166) show that there is a tendency for proximal terms to be learned first which seems to correspond to the various theories of markedness.

Fjalori i Gjuhës së Sotme Shqipe 'Dictionary of Today's Albanian Language' (AA. VV. 1980, from now on FGSS) defines as primary meaning of *atje* "the location far from us" which could be interpreted as far from both speaker and hearer (equivalent to Latin *illic*.) Its opposite is *këtu*. *Aty*'s primary meaning is defined as "the location far from me, but close to the interlocutor" or far from the speaker but closer to the hearer (Latin *istic*.) The opposite is

këtu for this word as well. *Këtu*'s primary meaning is defined as "the location close to us" referring to a location that is close to both speaker and hearer (Latin *hic*.) *Këtu* is provided with two opposites: *aty* and *atje*.

In the *Gramatika e Gjuhës Shqipe* 'Grammar of Albanian Language' (Agalliu et al. 2002, from now on GGS) there is not even an effort to define the meaning or functionality of these adverbs. The same indifference is noticed also in *Gramatikë Historike e Gjuhës Shqipe* 'Historical Grammar of the Albanian Language (Demiraj, 2002). The fact that grammars do not even try to deal with the meaning and functionality of the deictic adverbs of place appears to be a common phenomenon with other languages from the Balkan Sprachbund. For example, the Grammar of Literary Macedonian Language (Koneski 1981) only has a partial list of seventeen adverbs of place. The same was true for several Bulgarian, Romanian, Serbian, and Greek grammars leaving the whole task to lexicographical (FGSS) or etymological (Orel 1998) approaches.

From the etymological point of view, both *aty* and *atje* form a compound with the distal deictic prefix *a-*. The second elements *–ty* also found in some dialects as *–tu* and *–tje* are related to the old Indo-European pronoun *to-*. On the other hand, *këtu* has the same pronominal root which is preceded by the proximal prefix *kë-*.

The definitions offered by FGSS for the Albanian deictic locative adverbs assume a direct correlation between person and these adverbs suggesting that the authors of the dictionary define the Albanian deictic system primarily as a person based system. At the same time each definition includes also distance terms that would suggest some kind of hybrid deictic system involving both distance and person.

Data analysis

Based on the definitions above, the starting hypothesis is that *këtu* 'hic' should co-occur often with first person expressions, *aty* 'istic' with second person expressions and correspondingly *atje* 'illic' with third person expressions. First, second and third person expressions include either pronouns or verbs, singular or plural, in the corresponding person. Since *këtu* means "close to me," it is imaginable that, around it, there will be often words like *unë* 'I,' *ne* 'we,' *jam* 'be 1st singular,' *punojmë* 'work 1st plural,' etc. The same logic could be applied for *aty* and *atje*.

The following table contains the twenty highest ranked collocations of *këtu*. As discussed earlier, sorting is done by averaging the ranks of all three scores: Mutual Information, T-score and Z-score.

collocate	MI	t-score	z-score	m-rank	t-rank	z-rank	
përfshirë	8.11	17.23	101.29	16	1	1	6
përfshi	9.21	9.15	78.84	5	14	2	7
erdha	8.89	6.06	46.79	8	33	3	15
mbaron	8.31	6.01	37.91	14	34	7	18
jam	6.30	12.32	38.66	66	2	6	25
vij	8.02	4.94	28.16	20	49	11	27
kujtojmë	7.86	4.92	26.58	22	50	13	28
jemi	6.11	11.25	33.05	74	6	8	29
përmendur	6.92	6.13	23.86	42	31	18	30
pikërisht	6.05	10.63	30.70	78	8	9	32
përmend	7.16	5.26	22.23	35	40	23	33
llogaritur	7.67	4.61	23.21	25	56	20	34
bëjmë	6.32	7.31	23.12	63	24	21	36
ardhur	5.85	10.12	27.24	88	11	12	37
eja	8.54	3.66	25.00	10	92	14	39
ndodhem	9.22	3.42	29.59	4	105	10	40
shto	10.59	3.30	45.76	2	117	4	41
përmendet	7.44	4.27	19.85	31	67	26	41
rri	6.93	4.67	18.20	41	52	31	41
erdhëm	8.58	3.39	23.42	9	107	19	45

Table 13: Twenty highest collocations of *këtu*

The following table contains the English gloss and a grammatical tag that is also used in further analysis. The tags used for marking the grammatical tense follow Bertinetto's (1986) naming schema which covers well the Albanian indicative tenses. Since collocations are between single words, the compound tense counts are construed through manual analysis of concordances. Imperfect and Simple Perfect are easy to spot as they are synthetic forms. The only trace of the Compound Perfect that can be observed in the collocation table is a verbal adjective called the participle which is also used to form pluperfect, gerund, infinitive, and negative. To come up with a reliable number, three verbs (*kërkoj* 'search/ask,' *shkoj* 'go,' *shikoj* 'look') with varying frequencies were looked up for bigrams that disambiguate the participle as in the table below.

		kërkuar	shkuar	shikuar		
Non-finite	të	1400	845	16		
	duke	666	133	6		
	pa	33	21	0		
		2099	999	22	3120	24%
Compound Past	kam	209	107	2		
	ke	704	150	7		
	ka	3145	1072	3		
	kemi	213	51	2		
	keni	49	17	0		
	kanë	1433	745	2		
		5753	2142	16	7911	61%
	kisha	11	44	1		
	kishe	0	0	0		
	kishte	351	397	0		
	kishim	13	19	0		
	kishit	2	5	0		
	kishin	115	175	0		
		492	640	1	1133	9%

Table 14: Distribution of the participle among tenses and non-finite forms

Analytic non-finite forms make up 24% of the participle usage, Compound Perfect and pluperfect 70% and other forms (such as the articles *i*, *e*, *u*, *ju*, *ua*, the verb *është* 'is,' etc.) 6%. The numbers of the analytic pasts used from now on will reflect these percentages.

Collocate	English	Grammatical Tag
përfshirë	include	participle
përfshi	include	participle (Gheg)
erdha	come	1st simple perfect singular
mbaron	end	3rd present singular
jam	be	1st present singular
vij	come	1st present singular
kujtojmë	remember	1st present plural
jemi	be	1st present plural
përmendur	mention	participle
pikërisht	exactly	adverb
përmend	mention	1st present singular
llogaritur	calculate	participle
bëjmë	make	1st present plural
ardhur	come	participle
eja	come	2nd present singular
ndodhem	be situated	1st present singular
shto	add	2nd present singular
përmendet	mention	3rd present singular
rri	stay	1st/2nd/3rd present singular
erdhëm	come	1st simple perfect plural

Table 15: Translation and grammatical tags of *këtu* collocations

Table 15 offers a quick glimpse into the kind of words *këtu* attracts. Among the twenty top collocations 50 percent are 1st person tags, and 60 percent contain also a tag for the Present tense. In comparison, there are only 10 percent 3rd person tags and 20 percent Past tense tags. Upon further analysis, the Past tense tags, they all mark Perfect tenses (both Simple and Compound) and there are no tags for Imperfect.

The following tables replicate Table 13 and Table 15 for *aty* and *atje* respectively.

collocate	MI-score	T-score	Z-score	MI-rank	T-rank	Z-rank	
aty	7.37	18.28	83.02	8	1	1	3
ndodheshin	7.41	7.07	32.63	7	9	2	6
ndodhej	6.93	7.59	29.70	14	7	3	8
diku	6.76	5.93	21.80	19	16	7	14
mesi	8.27	4.36	27.07	3	37	4	15
orës	6.14	6.75	20.05	40	12	8	20
rri	6.82	4.15	15.59	17	39	12	23
pikërisht	5.82	8.88	23.65	61	5	5	24
jetojnë	6.08	5.29	15.38	45	26	14	28
mbërritur	5.90	5.68	15.51	53	20	13	29

rastësisht	6.59	3.78	13.14	23	47	18	29
isha	5.71	5.49	14.03	68	22	15	35
lamë	7.31	3.00	13.39	9	88	17	38
ora	5.78	4.68	12.26	63	33	22	39
shkon	5.79	4.60	12.09	62	34	24	40
armiku	6.75	3.21	11.76	20	77	25	41
gjeta	6.83	3.08	11.64	16	87	27	43
ndodhet	5.43	5.22	12.12	86	27	23	45
banojnë	6.89	2.95	11.34	15	93	29	46
gjetur	5.24	5.95	12.91	108	15	20	48

Table 16: Twenty highest collocations of *aty*

Collocate	English	Grammatical Tag
aty	there (istic)	adverb
ndodheshin	be situated	3rd imperfect plural
ndodhej	be situated	3rd imperfect singular
diku	somewhere	adverb
mesi	middle	noun nominative
orës	hour	noun genitive
rri	stay	1st/2nd/3rd present singular
pikërisht	exactly	adverb
jetojnë	live	3rd present plural
mbërritur	arrive	participle
rastësisht	accidentally	adverb
isha	be	1st imperfect singular
lamë	leave	1st simple perfect plural
ora	hour	noun nominative
shkon	go	3rd present singular
armiku	enemy	noun nominative
gjeta	find	1st simple perfect singular
ndodhet	locate	3rd present singular
banojnë	reside	3rd present plural
gjetur	find	participle

Table 17: Translation and grammatical tags of *aty* collocations

The top twenty collocations of *aty* include 25 percent Present and 30 percent Past tenses (three tags for Imperfect, two for Simple Perfect, and two participles.) More than half of the tags are for 3rd person expressions (marked as 3rd person verbs and nouns.)

collocate	MI-score	T-score	Z-score	MI-rank	T-rank	Z-rank	
shkova	8.10	5.83	34.12	13	11	2	9
shkuam	8.34	4.57	29.14	9	31	3	14
shkuar	6.47	8.63	28.68	43	1	4	16
shkonte	6.86	5.89	22.48	32	10	8	17
shkoja	7.56	4.17	20.31	18	40	10	23

shkojë	6.75	4.90	17.95	36	23	13	24
jetojnë	6.38	5.35	17.27	48	17	15	27
priste	6.86	4.27	16.24	33	37	19	30
atje	5.74	7.07	18.30	76	4	12	31
shkojmë	7.23	3.79	16.40	25	53	17	32
mbërritëm	8.67	3.10	22.13	6	86	9	34
shkuan	7.15	3.78	15.92	27	54	20	34
shkojnë	6.24	4.90	15.04	53	24	24	34
mbërrita	9.11	2.96	24.58	4	96	5	35
shkonim	7.58	3.32	16.24	17	70	18	35
mbërritur	5.99	5.24	14.76	64	20	25	36
qëndruar	6.21	4.63	14.07	56	30	28	38
arrihet	6.60	4.00	13.95	40	44	30	38
gjer	6.82	3.72	13.97	34	56	29	40
shkosh	8.14	2.92	17.33	12	102	14	43

Table 18: Twenty highest collocations of *atje*

Collocate	English	Grammatical Tag
shkova	go	1st simple perfect singular
shkuam	go	1st simple perfect plural
shkuar	go	participle
shkonte	go	3rd imperfect singular
shkoja	go	1st imperfect singular
shkojë	go	3rd present singular subjunctive
jetojnë	live	3rd present plural
priste	wait	3rd imperfect singular
atje	there (illic)	adverb
shkojmë	go	1st present plural
mbërritëm	arrive	1st simple perfect plural
shkuan	go	3rd simple perfect plural
shkojnë	go	3rd present plural
mbërrita	arrive	1st simple perfect singular
shkonim	go	3rd imperfect plural
mbërritur	arrive	participle
qëndruar	stay	participle
arrihet	arrive	3rd present singular
gjer	until	preposition
shkosh	go	2nd present singular subjunctive

Table 19: Translation and grammatical tags of *atje* collocations

What strikes the eye in this third set is the overwhelming presence of the verb *shkoj* 'go' in almost every person and in several tenses. There is an apparent imbalance of past and present tenses (2-to-1) with four tags for Imperfect, five tags for Simple Perfect, three participles and six Present tense tags. Meanwhile, an apparent balance is observed in the

distribution of the person: six 1st person tags, one 2nd person and eight 3rd person tags, or, a 7-to-8 ratio between person and non-person.

Grammatical features collocations

By rotating these tables around the grammatical tags, some interesting patterns emerge. The numbers are occurrences in the one hundred highest ranked collocations. These patterns hold for other groups of one hundred (100-200, 300-400) so this numbers can be considered percentages.

	1st person	2nd person	3rd person	discourse	present	perfect	imperfect
këtu	39	10	17	9	47	18	2
aty	8	2	46	2	18	22	27
atje	27	6	34	3	24	22	22

Table 20: Person and tense distribution in the one hundred highest ranked collocations

The data assembled in Table 20 shows that collocations for *këtu* are dominated by tags that refer to discursive situations. The presence of 39 first person tags is the only fact supporting the starting hypothesis that *këtu* should be surrounded by first person expressions. If the analysis is expanded to second person and complemented by discourse linguistic markers such as *ja*, *o*, *bre* an even more overwhelming 58% of the tags vs. the 17% third person tags, points to the discourse situation as being the main gravitational pull for *këtu*.

This is reinforced by the tense tags distribution: 47% are Present tense while Perfect and Imperfect constitute 18% and 2% respectively. This collocation of *këtu* and Present tense gives more weight to the fact that it occurs mainly in an act of speech. As for the past tenses, it is well documented that past tenses are used more in narrative or in reported discourse at least. Sakita (2002) reports ratios of 10-to-1 in favor of past tenses for speaker

reported discourse. The opposite ratios in our collocations lead to the conclusion that *këtu* favors direct discourse over reported discourse.

The data for *aty* is almost the opposite of *këtu*. The combination of first and second person tags as well as the discourse markers hardly reaches 12% of the collocations vs. the 46% of tags marking the third person. Past tense tags also make up 49% of the tags vs. the 18% marking the present.

As for *atje*, the data in Table 20 shows evenness of distribution among all its elements.

The presence of first and second person tags points to discourse relating to the speaker. The presence of present tense points to discourse relating to the time of speech. All the conclusions above can be collapsed in the following table:

	këtu	aty	atje
speaker involvement	+	-	ø
time of speech	+	-	ø

Table 21: Discourse features

Table 21 shows the picture of a triple opposition where *këtu* is opposed to *aty* and both of them are opposed to *atje* and very nicely exemplified in the graph below:

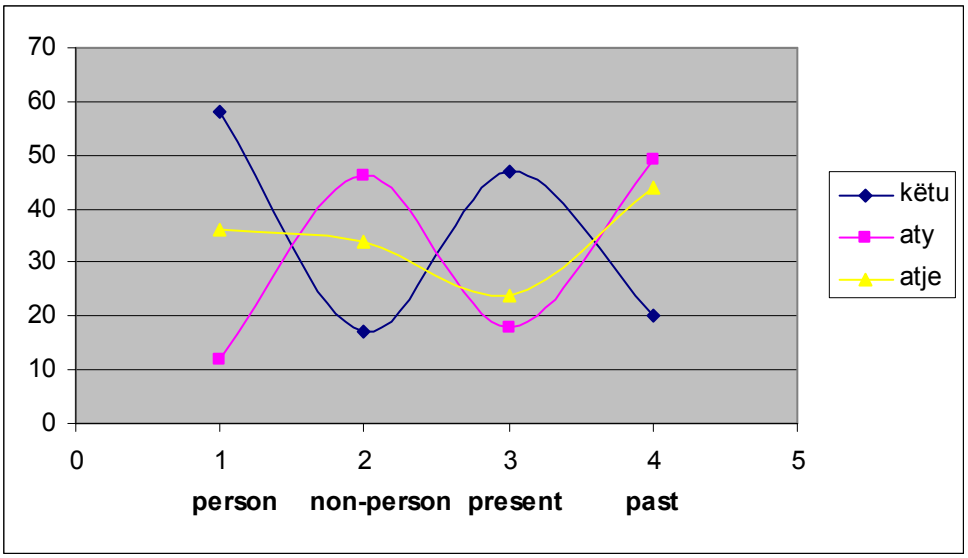


Figure 11: Distribution graph of features

It is interesting to note that while the graphs for *këtu* and *aty* follow two alternating lines, *atje* cuts more or less in the middle following a flatter line.

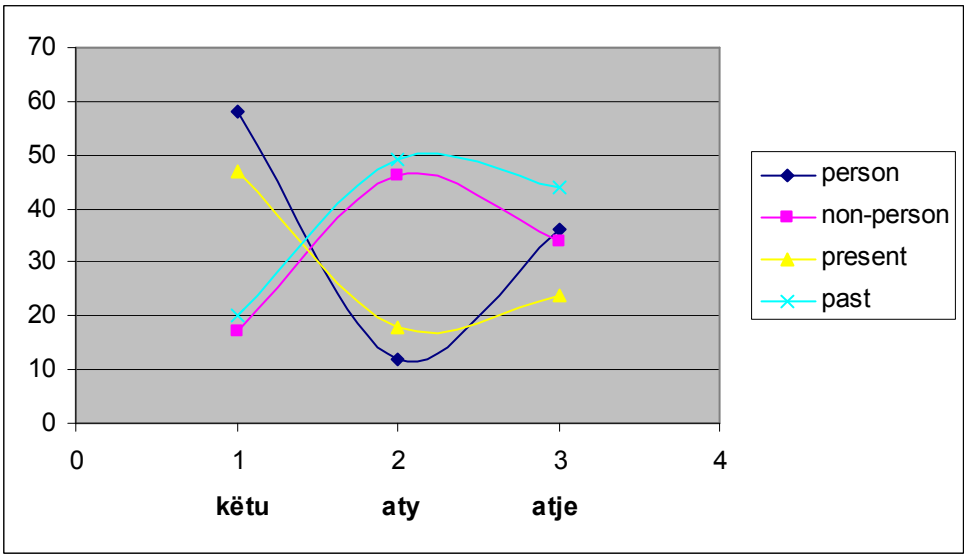


Figure 12: Feature divergence

Figure 12, which uses the same data, shows almost perfect pairing of feature collocations with the deictic adverbs of place: speaker participation follows the same line as present tense while the non-person (Benveniste 1966) line takes the same shape as the past tense. Person and tense features diverge for *këtu* and *aty* and converge for *atje*.

The conclusion of this part of the data analysis can be illustrated by the following figure:

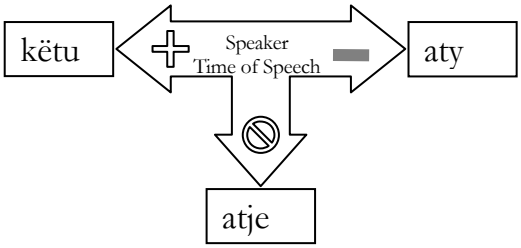


Figure 13: Opposing features - opposite words

Both *këtu* and *aty* have features whose distribution differentiates them from each other. Meanwhile *atje*, having an equal distribution of most of the analyzed grammatical features, either does not belong in this group forming a different paradigm or it represents an alternative to the first two. The list of the analyzed features and their respective distribution among these three adverbs does not suggest a connection to the person but a rather strong correlation with the moment of speech.

Semantic collocations

Using a different lens in the observation of the collocations lists generated by the three adverbs of place, other patterns emerge. The first observation is that the largest category of words collocated with these adverbs are verbs: 68, 67, and 69 percent respectively for *këtu*, *aty* and *atje*. A word is considered a top collocation if it is listed below a cutoff value which depends on a compounded score and that more or less corresponds to the first two hundred words for *këtu* and the first one hundred for *aty* and *atje*. The word *këtu* reaches the cutoff value at a higher rank because of its larger number of collocates as well as the larger number of senses.

Collocations with verbs

There are certain verbs that collocate with all three deictic adverbs and they are represented in the table below:

	këtu				aty				atje			
	#	f	F	ratio	#	f	F	ratio	#	f	F	ratio
arrij 'to arrive'	1	40	5604	140.10	3	81	7811	96.43	1	19	804	42.32
gjej 'to find'	1	6	306	51.00	5	**105**	7455	71.00	3	52	6114	117.58
jam 'to be'	*7*	*438*	*20234*	*46.20*	*4*	*190*	*20283*	*106.75*	*5*	*172*	*26962*	*156.76*
jetoj 'to live'	4	51	3105	60.88	3	69	5157	74.74	4	66	3875	58.71
kam 'to have'	*4*	*430*	*37428*	*87.04*	*1*	*195*	*35770*	*183.44*	*4*	*139*	*22630*	*162.81*
kujtoj 'to remember	2	44	641	14.57	2	28	1865	66.61	1	6	34	5.67
mbërrij 'to arrive'	1	8	456	57.00	2	53	2793	52.70	3	55	2505	45.55
ndodhem 'to be located'	1	12	56	4.67	4	**185**	7463	40.34	4	52	4622	88.88
punoj 'to work'	2	28	2307	82.39	1	9	491	54.56	5	88	6139	69.76
qëndroj 'to stand'	4	40	1956	48.90	6	82	4972	60.63	5	63	3837	60.90
rri 'to stay'	4	51	1033	20.25	3	38	1268	33.37	2	20	956	47.80
shkoj 'to go'	1	6	156	26.00	4	95	7593	79.93	15	**373**	14404	38.62
vij 'to come'	14	**436**	18354	42.10	1	22	2195	99.77	1	6	91	15.17

Table 22: Verbs that collocate with all three adverbs

Table 22 can be analyzed following the three highlighted sections, one for each adverb, where each row corresponds to the lexical headword[20]. The first column of each section contains the number of different forms (e.g. *jetoj* 'to live' has four forms collocating with *këtu*: *jetuar*, *jetoj*, *jetojmë*, *jetojnë*) that appear in the top collocations. The second column (marked by *f*) contains the number of collocations of these word forms with the corresponding adverb. The third column (marked by *F*) contains the total number of

[20] In Albanian lexicography, the verbal headword is the first person, singular, present, indicative form of the verb.

occurrences of these forms[21] in CALT. The final column contains the ratio of the word collocated with the corresponding adverb against the total number. The ratio should be read as 1-to-N where N is the number displayed in the respective column. The low ratios correspond with high MI scores[22] illustrating once again one of the weaknesses of the MI computation.

For some words, the number of top collocations is evenly spread among columns. An explanation could be found in the common traits adverbs of place have: they all attract verbs that contain some kind 'stationing' in their semantic definition such as *jetoj* 'to live,' *qëndroj* 'to stand,' and *rri* 'to stay.' The number of forms for the verb *arrij* 'to arrive' is evenly collocated with *këtu* and *aty*. This distribution and the fact that collocates with *atje* are much lower, contributes to a semantic feature not found in FGSS: 'get to a defined place without regard to the direction of movement.'

The most uneven distribution of a collocated verb with the three adverbs of place is the deictic verb *vij* 'to come' collocating 436 times with *këtu* and only 22 and 6 times with *aty* and *atje* respectively. The verb *vij* 'to come' points to a well defined reachable location.

---

[21] Since these numbers represent the total occurrences of possibly different sets of forms they can have different values. For example, *aty* and *këtu* have both a count of two word forms for *kujtoj* 'remember' collocated with them. The two forms of *kujtoj* for *këtu* are *kujtoj* and *kujtojmë* respectively 329 and 312. For *aty*, the two forms are *kujtohet* and *kujtuar* respectively 1101 and 764.

[22] For example, the verb *arrij* has a ratio of 1-to-140 in the *këtu* column, 1-to-96 in the *aty* column, and 1-to-42 in the *atje* column. Their respective MI scores are 4.32, 5.46, and 6.60.

Very uneven distributions (or meaning defining collocations) pertain to the verb *shkoj* 'to go' which weighs heavily in the meaning of the adverb *atje* 'there.' This adverb appears as a semantically monolithic word whose usage can be defined by the 373 occurrences with the verb *shkoj*. Vice-versa, the verb *shkoj* defines itself through its own list of collocations that is dominated by *atje* 'there,' *larg* 'far,' *tej* 'beyond,' *ngado* 'wherever,' *tutje* 'farther/yonder,' and *kudo* 'anywhere' as in Table 23.

shkoj	MI	T	Z
atje	5.71	11.40	29.15
larg	5.42	8.53	19.72
tej	5.35	9.34	21.12
ngado	6.82	2.79	10.47
tutje	5.58	3.90	9.53
kudo	5.32	4.45	9.94
nga	2.10	-21.00	-15.44

Table 23: Collocations of the verb *shkoj* 'to go'

The verb *shkoj* 'to go' points to a loosely defined, maybe invisible location.

The other differentiating collocation in Table 22 is the verb *vij* 'to come' that represents 14 different forms all found in the list of top collocations for *këtu* 'here.' This verb defines *këtu* the same way *shkoj* 'go' defines *atje*. The verb *vij* is further defined by the collocations it creates on its own.

vij	MI	T	Z
nga	3.39	11.80	13.56
këtu	4.30	8.20	12.87
aty	3.94	5.40	7.47
larg	3.75	3.16	4.10
atje	3.09	0.44	0.46
kudo	2.55	-1.05	-0.89
tej	2.67	-1.58	-1.41

Table 24: Collocations of the verb *vij* 'to come'

One of the top collocations of the verb *vij* is the preposition *nga* 'from.' This same preposition has negative T and Z values for the verb *shkoj*. Since the preposition *nga* requires

a defined origin, by extension, the verb *vij* has the same semantic feature while the verb *shkoj* does not. The verb *vij* has no collocations at all with *ngado* 'wherever' and *tutje* 'farther/yonder.' It has negative T and Z values for *kudo* 'anywhere' and *tej* 'beyond.' At the same time, the collocation strength of the verb *vij* is high with *këtu* and *aty*, and close to zero with *atje*.

The other collocation specific to *këtu* is the verb *iki* 'to leave' with two forms and high level scores (MI: 6.40, T: 5.36, Z: 17.42). It is interesting to note that this verb is the opposite of *vij* and that it relates to *këtu* because of the meaning of directionality: to and from the *origo* as defined by Bühler (1934/1990:117) and part of which *këtu* is. Therefore, *këtu* collocates with all three elements of the *origo*: place of speech, moment of speech, and speaker. The speaker, represented mostly by first person verbs, is also marked through the use of the personal pronoun *unë* 'I.' Even though Albanian does not need the explicit use of the first person pronoun (Murzaku, 1988), its collocation strength (MI=4.97, T=11.47, Z=22.75) with *këtu* ranks 39th out of 1034 as opposed to 149th out of 850 for *aty* and 65th out of 759 for *atje*.

The other unbalanced distribution belongs to *aty* which has the verbs *ndodhem* 'to be located' and *gjej* 'to find' as its dominant collocates. These two verbs add to *aty* a meaning of visible, well known, and static location of an object. In Table 25, *gjendem* 'to be found/to be located' is an exclusive collocation of the adverb *aty* and, since the meaning of this word is very close to *gjej* 'to find/to be found,' the two representatives of ten different forms define the adverb *aty* as different from both *atje* and *këtu*. These three verbs (*gjej*, *gjendem*, and *ndodhem*) share the sense of 'finding/being found' or 'locating/being located' which in turn contain a sense of a definable place (as opposed to *atje*) and that it is not intrinsically in a

defined place (as opposed to *këtu*). The verb *banoj* 'to reside' and *strehoj* 'to shelter' complete this list.

The differences between the three adverbs are also illustrated by the exclusive groups of collocations created by each of them in Table 25.

Verb	English	këtu				aty				atje			
		#	f	F	ratio	#	f	F	ratio	#	f	F	ratio
bëj	to make	3	154	13790	89.55								
fus	to insert	1	19	2026	106.63								
iki	to leave	2	24	1040	43.33								
jap	to give	1	10	594	59.40								
mungoj	to be absent	1	13	1037	79.77								
ndihem	to feel	1	7	248	35.43								
pëlqej	to like	1	15	745	49.67								
sëmurem	to get sick	1	6	78	13.00								
vdes	to die	1	9	636	70.67								
vlej	to be worth	1	15	775	51.67								
banoj	to reside	1	12	276	23.00	3	24	816	34.00				
lind	to be born	1	25	1960	78.40	1	14	693	49.50				
shoh	to see	2	28	1161	41.46	1	6	139	23.17				
dërgoj	to send	1	8	330	41.25					1	8	330	41.25
kthehem	to return	1	6	222	37.00					1	9	222	24.67
pres	to wait	2	14	570	40.71					1	21	746	35.52
shikoj	to look	1	7	296	42.29					1	7	296	42.29
afroj	to approach					1	6	43	7.17				
fle	to sleep					1	7	170	24.29				
fsheh	to hide					1	16	1226	76.63				
gjendem	to be found					4	37	1853	50.08				
hyj	to enter					1	12	903	75.25				
kaloj	to pass					2	28	2016	72.00				
lë	to leave					3	87	8602	98.87				
ndërtoj	to build					1	6	88	14.67				
njoftoj	to announce					1	6	193	32.17				
pi	to drink					1	7	160	22.86				
strehoj	to shelter					1	10	365	36.50				
ul	to lower					1	16	1467	91.69				
zhvillohem	to be developed					1	8	378	47.25				
mbetem	to be left					1	44	4356	99.00	1	12	1049	87.42
largohem	to leave									1	6	34	5.67
mbij	to sprout									1	6	11	1.83
nis	to begin									1	7	274	39.14
shkel	to step on									1	6	287	47.83
shtroj	to lay down									1	13	953	73.31
studioj	to study									1	6	49	8.17
thërras	to call									1	7	128	18.29

Table 25: Distribution of verb collocations among adverbs.

Another group of semantically similar verbs form a separate subset from the rest of the collocations for *këtu*. They are listed in the table below.

Verb	English	këtu				aty				atje			
		#	f	F	ratio	#	f	F	ratio	#	f	F	ratio
dua	to want	3	84	4068	48.43					1	12	1154	96.17
flas	to speak	2	24	1091	45.46								
harroj	to forget	2	28	1502	53.64	1	7	110	15.71				
kujtoj	to remember	2	44	641	14.57	2	28	1865	66.61				
kuptoj	to understand	1	13	292	22.46								
llogaris	to calculate	1	23	317	13.78								
mbaroj	to complete	4	73	1962	26.88								
mendoj	to think	1	36	3979	110.53	1	9	338	37.56				
merrem	to deal with	1	6	211	35.17								
ndalem	to be stopped	3	24	341	14.21								
përfshij	to include	4	423	4407	10.42								
përfundoj	to conclude	1	19	975	51.32								
përgjigjem	to answer	1	13	1036	79.69								
përjashtoj	to exclude	3	37	1008	27.24								
përmend	to mention	5	114	2324	20.39								
shtoj	to add	4	60	2594	43.23								
sjell	to bring	2	25	1308	52.32								
them	to say	4	90	7723	85.81					2	25	2041	81.64

Table 26: List of collocations of verbs referring to speech.

This list comprises 44 forms that are a large percentage (41%) of the top verb collocations of *këtu*, implying therefore an additional meaning for this adverb. Each of these verbs refers to the act of speaking or to the speech content itself. The following list of concordances is an illustration of this particular use, missing in both FGSS (1980) and the newer FSS (1984).

pa harruar	këtu	,edhe debatin mes Berishës dhe Gjinushit
Ai përmendi	këtu	edhe projekte të kompanisë tjetër
çdo republikë, duke përfshirë	këtu	edhe Kosovën
duke shtuar	këtu	se në atë kohë u tha se
pa llogaritur	këtu	me qindra koncerte
Le të kujtojmë	këtu	vetëm titujt e tragjedive

Concordance 1: Anaphoric uses of *këtu*

In each of these expressions ('without forgetting here,' 'he mentioned here,' 'including here,' 'adding here,' 'without calculating here,' 'let us remember here'), the adverb *këtu* is not referring at all to a place but rather to the discourse itself. Referring to a discourse entity, gives the adverb *këtu* an obvious anaphoric role.

Collocations with other parts of speech

An easy collocation to spot are the numeric values such as 1, 2, 3, 2000, 2001, 2002, that can be counted by just searching for numeric values. The three adverbs return the following counts: *këtu* - 22; *aty* - 34; *atje* - 10. In addition to the numbers already counted, a large subset of collocates for *aty* is formed by a group of words expressing numbers or time like *ora* 'hour,' *mesi* 'middle,' *fundi* 'end,' and *viti* 'year' as shown later in Table 31. The obvious higher collocation of this kind of words points to another accepted meaning for *aty* and *këtu*: that of time. The following concordances illustrate this conclusion.

një herë	aty	nga viti 1970 dhe për herë të dytë në vitin 1976.
ku ju internuan së fundi	aty	rreth vitit 1983
një ditë të bukur doli nga burgu,	aty	rreth viteve 1988,
shumë aktore të bukura vetëvriten	aty	rreth të 40-tave a të 45-tave
autorëve më të lashtë të botës qysh	aty	e dymijë e pesëqind vjet më parë,
Do të ishte dashur të zhvillohej	këtu	e një vit të shkuar.
i ka falur çështjes tërë rininë dhe tërë mëndjen	këtu	e njëzet e shtatë vjet me radhë

Concordance 2: Time references using *aty* and *këtu*

The examples in Concordance 2 show references of time such as 'there by year 1970,' 'there around the year 1983,' 'there around the years 1988…,' 'there around the age of 40 or 50,' 'there and 2500 years earlier,' 'here and one year past,' 'here and 27 years continuously.' These collocations point to and confirm a second meaning offered by both FSS and FGSS defined as "approximately at that time." Some of these expressions are found with *këtu* as well confirming one of the secondary meanings found in FSS "from this time" as in the expression *këtu e dy vjet* 'from now to two years ago/later .' The difference between the two appears to be topologic: *aty* points to an approximate time line open at both ends while *këtu* points to a well defined segment of time with a clear start and end. The expression *këtu e [amount of time]* has also a variant with *aty* like in *aty e dy vjet* 'from than to two years earlier/later.' It is interesting to notice in this differentiation that the expression with *aty*

points to a time segment following or preceding a reported moment (most likely in the past) while the expression with *këtu* points to a time segment following or preceding the moment of speech. This are better labeled as anaphoric (*aty e* …) and deictic (*këtu e* …) temporal references as explained by Bertinetto (1986:30). Even though there is a small set of similar expressions collocated with *atje* there were not any similar constructs in any of its 4615 concordances. The same lack of collocations between *atje* and temporal expressions can be seen in Table 29 where the values for *dikur* 'sometime in the past' are zero [= no collocation]. This new feature of pointing to time reaffirms the separation of *atje* from *këtu*/*aty* as seen in the following table:

	këtu	aty	atje
closed topology	+	-	ø
deictic reference	+	-	ø

Table 27: References to time

From the analysis of the collocation data emerge some other words that point to semantic differences between *këtu*, *aty* and *atje*. The first one is the modifier adverb *pikërisht* 'precisely.'

pikërisht	MI-score	T-score	Z-score
këtu	6.06	10.63	30.70
aty	5.82	8.88	23.65
atje	4.82	4.64	8.72

Table 28: Collocation scores of *pikërisht*

All three adverbs have *pikërisht* as one of their higher collocations. Their scores are in descending order from *këtu* to *aty* to *atje*. These scores become more meaningful when added to the discussion of time references. The word *pikërisht* 'precisely' is more likely to appear near *këtu* which also has a feature of boundedness in its time references. The probability for *pikërisht* to be near one of these adverbs decreases in parallel with the decrease of boundedness in *aty* and its disappearance in *atje*. These parallels could be drawn even further

when compared to the grammatical tense and person in which *këtu* attracted first person and Present, *aty* third person and Past and *atje* every person and tense.

The collocation strength with words such as *diku* 'somewhere' and *dikur* 'sometime in the past' shows again that *këtu* has difficulty to relate to unbounded expressions in opposition to *aty*'s numbers shown in the table below.

	diku			dikur		
	MI-score	T-score	Z-score	MI-score	T-score	Z-score
aty	6.76	5.93	21.80	5.40	3.14	7.22
atje	5.33	2.77	6.22	0	0	0
këtu	4.36	1.83	2.92	3.87	1.11	1.50

Table 29: Collocation scores with *diku* and *dikur*

Since *atje* does not relate to time concepts, it has no collocations at all with the word *dikur* 'sometime.'

One of the most widely accepted uses of the adverbs of place is its anaphoric use as pro-forms for locative prepositional phrases. The most common locative preposition is *në* 'in/at/on' which is also the third most frequent word in CALT at 34 per thousand. Because of this high frequency it is expected that T and Z scores which are based in deviation from the norm would be negative as they are in the table below.

në	MI-score	T-score	Z-score
këtu	2.05	-29.15	-21.32
aty	2.34	-18.52	-15.01
atje	2.52	-11.94	-10.29

Table 30: Collocation scores of *në*

There is a gradual increase of the MI score while the deviation decreases from *këtu* to *aty* and then to *atje*. The attraction of *atje* for *në* infers that *atje* needs a further definition with a locative expression. The forms *aty* and *këtu* need the reinforcement less. Therefore *këtu* is the better defined location, followed by *aty* and trailed by *atje*.

Since there is an obvious connection of the kind of deixis and tense, we had hoped to find some connection to the corresponding semantic feature, telicity. Bertinetto (1986) runs several tests by checking, among other things, the compatibility of adverbs with verbs of varying telic values. Di Giovine (1990:21), in his comparative analysis of the perfect tense formation in Indo-European, integrates the study of the relations between verbal semantics and the presence or absence of perfect tense in the object languages. By populating the list of collocated verbs with actionality features extracted from Bertinetto (1986, 2001), no clear attraction of one kind or another was found. The three features used to define the actionality status (statives, activities, achievements, or accomplishments) are ±durative, ±dynamic, and ±homogeneous (see Appendix: Verb Actionality.)

The distribution of nouns is represented in the table below. The adverb *aty*, in addition to the numeric values which have already been discussed, collocates with a large number (15%) of expressions of time (*dita* 'day,' *ora* 'hour,' *viti* 'year,' etc.) These being definite forms, the idea that *aty* refers to a well known entity is reinforced.

	këtu		aty		atje	
location	64	20%	69	25%	57	27%
time	23	7%	40	15%	17	8%
numeric	22	7%	34	12%	10	5%
human	79	25%	67	24%	50	24%
other	133	41%	65	24%	78	37%
total	321	100%	275	100%	212	100%

Table 31: Collocated nouns

A further subcategorization not represented in Table 31 and that differentiates the three adverbs is observed in nouns of location. Both *këtu* and *atje* collocate with geographic proper nouns such as *Amerikë*, *Shqipëri*, *Tiranë*, *Itali*, etc. The word *aty,* however, has a very insignificant number of proper nouns and all of them at the bottom of the list with negative

T- and Z-scores. What dominates its list of collocates are more familiar places such as the following top ten collocations in order: *shtëpi* 'house/home,' *dyqan* 'store,' *hotel* 'hotel,' *tavolinë* 'table,' *bar* 'bar,' *varr* 'grave,' *tokë* 'land,' *derë* 'door,' *pallat* '(apartment) building,' *lokal* 'bar/restaurant.'

The strength of the noun collocations is interesting as well. The strongest twenty percent of the noun collocations for *këtu* include in order only *Amerikë*, *Shqipëri*, and *Tiranë*. The majority of the stronger collocations is made of mostly abstract nouns such as *fjalë* 'words,' *udhëzimet* 'the instructions,' *puna* 'the work,' *çudia* 'the wonder,' *problemi* 'the problem.'

The list of the ten strongest noun collocations for *aty* mentioned above is all included in the top twenty percent.

The strongest collocations for *atje* include the country names *BRSS* 'USSR,' *Amerikë*, *Itali*, *Greqi*, *Gjermani*. The other nouns included in the top twenty percent have also to do with concepts associated with 'far' and 'away' such as *emigrantëve* 'emigrants DAT,' *pushime* 'vacations,' *ambasadës* 'embassy DAT.' These data offer a new view: *këtu* is self sufficient in describing a location which is usually the *origo*. Once this word is uttered, the speaker's location is known and there is no need to name it too. This is not the case with *aty*. This word refers to nearby locations that need to be identified in the deictic field. As for *atje*, it refers to undefined places located outside the deictic field. These places need to be identified and what better device than a proper noun.

The fourth deictic adverb

In their map of the two- and three-term deictic systems in the Mediterranean, Benedetti and Ricca (2002) point to an exception by the deictic system of Arbëresh - a

dialect of Albanian - that contains four terms. Complex deictic systems are found in several areas of the world such as the Northeast Caucasus where Lak, with its three *there's* (Friedman 1994), has a five term deictic system. Having two deictic systems in the same language might shed some light on how the geometry of deictic terms evolves. The double opposition described up to this point shows traces of a fourth deictic.

The words belonging to deictic sets in Albanian start with either *a-* or *kë-* and this regularity expands across several parts of speech such as pronouns *ai/ky* 'that/this NOM,' *atij/këtij* 'that/this DAT,' adverbs *ashtu/kështu* 'thus,' adjectives: *i atillë/i këtillë* 'such,' etc. These prefixes are not productive anymore but their compositional elements are noticeable and the derivation through affixation is transparent. Following the same pattern, *kë-* is a prefix in *këtu* and *a-* in *aty*. Even though apparently *aty* has a differently realized vowel in the end, these two words share the root *–tu* as Çabej argues in his etymological analysis (1976a:282). In fact, there are several dialects that still today use the form *atu*. Meanwhile, the root *–tje* of *atje* is also found in the word *tutje* which Çabej (1976b:198-199) analyzes as the compound of the two primitives *tu+tje* meaning initially 'from here to there' and later becoming simply 'far away.' The regularity of these binary oppositions would presume that even *atje* should have an equivalent in *kë-* and that this word should be *këtje*.

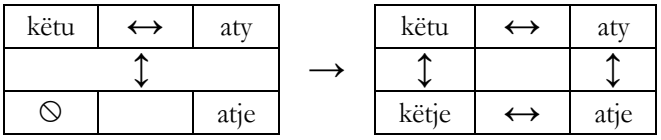


Figure 14: From three to four term opposition

This fourth deictic is not used anymore in standard Albanian, but it appears in writings of both Gheg and Tosk authors like Frang Bardhi (17th century), in poems of the "bejtexhinj" (18th century), Naim Frashëri (19th century), and Martin Camaj (20th century). In

the Arbëresh dialect continuum of southern Italy it is common in everyday use and it has been documented in several linguistic descriptions of the Arbëresh variety of Acquaformosa (Raimondo, 2001), the Crotone area (Turano, 2001), Falconara Albanese (Camaj, 1977; Altimari and De Rosa, 1995), San Costantino Albanese (Scutari, 2002), Vacarizzo Albanese (Hamp 1954/1993) as well as in De Rada's *Songs of Serafina Topia* (De Rosa, 2003).

Even with its widespread use in Arbëresh, defining *këtje* proves to be an elusive enterprise. During interviews with Arbëresh speakers, they had a hard time separating *atje* from *këtje* but they feel that *atje* is more distant. The mini-corpus built from the examples extracted in the texts mentioned above, contains twenty-two different usages of which seven have in the immediate vicinity the word *lart* 'up' or *mbi* 'over.' In one of them *këtje* refers to a house underground keeping with the vertical dimension.

In the following two examples every native speaker can understand but cannot explain the opposition between *atje* and *këtje*.

kanë vajtur deri atje e deri këtje
'they went up to there.A and up to there.K'
-Atje e kanë bukën e mirë. -Jo! Këtje e kanë më të mire.
'-There.A the bread is good. -No! There.K it's better.'

The other twelve are used in the fixed expression *këtu-këtje* meaning 'here and there.' The Albanian standard variant of this last expression is *aty-këtu* 'there and here.'

The dominating collocation of *lart* 'up' with *këtje* can be used as an argument for the existence of the vertical dimension in the Arbëresh deictic system such as in Lak (Friedman, 1994) or in Belhare (Plank, 1998:117). Since the obtained concordances contain only 715 words, there is not enough evidence to justify such a conclusion. Instead, the collocation with *lart* as well as the analysis of the contexts where the word was found would lead to a

meaning of well defined faraway location such as the top of the mountain, a house that could be seen from the window of the school, and the house of the Turkish sultan. The same meaning could also be extracted from the generic advice *Ndë Shin Ndre bën valt e vëre këtje* 'By Saint Andrew's feast, make oil and put it there.K' where *këtje* marks a location that is generic to the speaker but should be well defined to the receiver of the advice.

The evolution of the deictic adverbs in Albanian

The group of the deictic adverbs is fluid, and words enter and leave the system with the introduction of new linguistic devices or infrequent use. Individual members of closed classes are lost, usually by one member generalizing to take over the functions of other members (Krug, 2001). Such an example is shown by Žic Fuchs (1996) in the standard spoken in the city of Zagreb in Croatia. Standard Croatian preserves today the three-degree spatial distinction based on participant roles taken from the Stokavian dialect. Unlike the rural variant, Croatian spoken in the city of Zagreb indicates that the medial distance category or 'proximal to the Hearer' is disappearing. The two deictic adverbs *ovdje* and *tu* have merged and simply mean 'proximal to the Speaker.' The same is happening in the three member set of sentential demonstratives *evo*, *eto*, *eno* but, in this case, the medial is merging with the distal demonstrative.

The different deictic sets in the archaic Albanian dialects show that this category is in flux and has been changing in the last five hundred years. According to Çabej (1976a:282) Albanian started with one deictic whose only meaning was "in a defined place" without any regard to speaker or distance. This primitive root is **tu* related to the Indo European **to*. It is also found in *tutje* 'over there,' *tëhu* 'over here' and *njitu* 'right here.' While one-degree deictic systems are rare, there is at least one other one-degree deictic adverb class reported by Frans

Plank for Bavarian in his Raritäten database (1998:51). Given this situation, we could propose this schematic of a possible evolution:

inventory of forms						occurrences in CALT
*to	*ko + *to	*ka + *tu	kë+tu	këtu	këtu	6789
	*ao + *to	*a + *tu	a+ty	aty	aty	5855
		*te	kë+tje	këtje		
			a+tje	atje	atje	4615

Table 32: Evolving geometry of the adverbs

The forms based on the primitive root **tu* have the longest continuous line. The word *këtu* being the oldest form has brought about the highest number of semantic syntactic functions (deictic proper, anaphoric, temporal) and therefore its highest frequency. It is followed by the next highest number of occurrences and the two functions of *aty* (deictic and temporal). The list finishes with *atje*'s one deictic function and its lowest number of occurrences.

Table 32 can be summarized in the following way: Proto-Albanian started with just one deictic adverb **to* which meant "in a defined place." This adverb was very often preceded by one of the deictic particles **ko* 'proximal' or **ao* 'distal.' By the end of the Proto-Albanian period was introduced the adverb **te* meaning "in an undefined place" which, according to Eric Hamp (p.c.), is related to Lithuanian **te*. By analogy, the two deictic particles were preceding this new adverb as well.

In the same way, some demonstratives formed the noun case endings and verbal inflection, these particles move from an analytic device into a synthetic deictic marker sometime after the Latin influence on Albanian had exerted its modifications. This perhaps explains the fact that *a-* was not reduced into an *ë-* or eliminated.

Sometime during the sixteenth century, the Arbëresh diaspora was created in Italy where the four terms of the deictic adverb system resisted to this day. In the main Albanian dialects, the functions of the forms *këtje* started being taken by *aty* which covered the undefined but visible location/reference. The word atje was left with its original undefined location/reference. The word *këtje* is not needed anymore.

Deictic directional adverbs *andej* and *këndej*

The total number of occurrences in CALT is 901 for *andej* 'on that side' and 233 for *këndej* 'on this side.'

The number of collocations generated by each of these words reflects the total distribution: there are 148 collocates for *andej* and 40 for *këndej*. The main collocates of the directional adverbs *andej* and *këndej* are complementary since their main collocation is with each-other forming the idiom *andej-këndej* 'all over.'

andej				këndej			
	MI-score	T-score	Z-score		MI-score	T-score	Z-score
				andej	12.76	8.93	263.53
herë	3.99	1.72	2.42	herë	5.16	2.05	4.35
këndej	11.45	5.73	107.27				
prej	7.52	15.92	76.37	prej	5.68	3.78	9.57
shpërndarë	8.17	3.07	18.45	shpërndarë	9.61	2.62	25.85
				tutje	9.44	2.42	22.53
vetëm	2.16	-2.24	-1.68	vetëm	4.11	1.52	2.23

Table 33: Prefab and idiom identifying collocations

Indeed, the highest collocations are formed between words that form prefabs and/or idioms. Besides the idiom *andej-këndej*, we find the prefabs *herë këndej, herë andej* 'here and there' and *shpërndarë andej-këndej* 'scattered all over.' The preposition *prej* 'from' is found in this table because of the ablative pronominal root **ndej*. The adverb *tutje* 'away' also forms the prefab *këndej e tutje* 'from now on.' In Table 33 is included the word *vetëm* 'only' even though it does not create an idiom with *andej* and *këndej*. This word, besides the 1-to-2 ratio of the MI score, contains opposing values in their T and Z scores: negative for *andej* and positive for *këndej*. The reason this word is included in the table above is its potential for exposing the meaning differential between these two deictic words. Expressions of specificity tend to gravitate around the proximal term. As *pikërisht* 'exactly' was seen to appear mostly with *këtu* 'here,' the pair *andej/këndej* is defined by *vetëm*, making *andej*

undetermined and *këndej* determined. The adverb *diku* 'somewhere' (MI=6.68, T=2.26, Z=8.10) appears exclusively in *andej*'s collocations contributing to its undeterminedness.

Another defining word, exclusive to *këndej*, is *kufirit* 'border / limit: singular dative / ablative.' The meaning of border / limit too contributes to the determinateness of *këndej* vs. the undetermined *andej*. The same analysis can be applied to the adjoining of *tutje* in the idiom *këndej e tutje* where *tutje* stands for unbounded infinity.

As the most generic of the pair, *andej* is more frequent and creates more collocations. Its collocations include several forms of the verbs *vij* 'to come,' and *kthehem* 'to return' of which six are past tense forms and two present. Also, there are several forms of the verbs *largohem* 'to get away,' *nxjerr* 'to take out,' *dal* 'to go outside,' and *shkoj* 'to go.' Of the eighteen forms these six verbs take, eleven are in past tenses and seven in present. Past tenses for both groups of verbs have higher collocation scores. Meaning wise, we notice that there is a greater variety of verbs depicting actions of movement away from the *origo* towards an undefined place marked by other high collocated words such as *larg* 'far,' *mal* 'mountain' and *rrugë* 'road.' The verbs depicting actions in the opposite direction (towards the *origo*), are justified by the highly collocated preposition *prej* 'from' as in …*kanë ardhur prej andej*… '...they came from there...'

Deictic adverbs of manner *ashtu* and *kështu*

The total number of occurrences in CALT is 9125 for *ashtu* 'thus / in that manner' and 8167 for *kështu* 'thus / in this manner.' Çabej (1976a) sees in these words the combination of the deictic part *a-* or *k(ë)-*, an interrogative pronoun *ç* and the root *–tu* of *aty/këtu*. These words are indeed pronouns that have become adverbs through a hypostasis process (Çabej 1986:240).

In the list of top collocations for both *ashtu* and *kështu* is the particle *po* which forms the prefab *po ashtu* / *po kështu* 'just that/this way.' The list continues with verbs which are somewhat different between the two: *ashtu* is collocated with five different forms of the verb 'to act' (*vepruar*, *veprohej*, *veprua*, *veproi*, *vepruan*), and two forms of the verb 'happens' (*ndodhi*, *ndodh*). Whereas *kështu* is collocated first with five forms of 'happens' (*ndodh*, *ndodhë*, *ndodhte*, *ndodhi*, *ndodhur*), and three forms of 'to act' (*vepruar*, *veprohet*, *vepronin*). This asymmetry is explained by the almost fixed expressions *ndodh kështu* 'it happens this way' and *ashtu siç veprohet* 'the same way it is acted / done.' While the former is an anaphora for some action just described, the latter is a similitude to the way things are done elsewhere as depicted by other collocations such as *Amerika* 'America,' *gjermanët* 'the Germans,' *bota* 'the world,' *lindja* 'the east,' *perëndimi* 'the west.'

The other collocations include verbs. The adverb *kështu* has 19 verbal forms or pronouns in the first or second person. *ashtu* instead has only one form in first person plural. As with the adverbs of place, the *k(ë)-* adverb appears more closely related to the act of speech, i.e. conversational style, while the *a-* adverb takes a more narrative functionality.

By observing the tenses of the verbs, we get an equal distribution of forms for *ashtu* represented by a flat line in Figure 15. This is not the case for the adverb *kështu* that creates two peaks (the highest one for present and the next one for the compound past, with a low point for imperfect.)

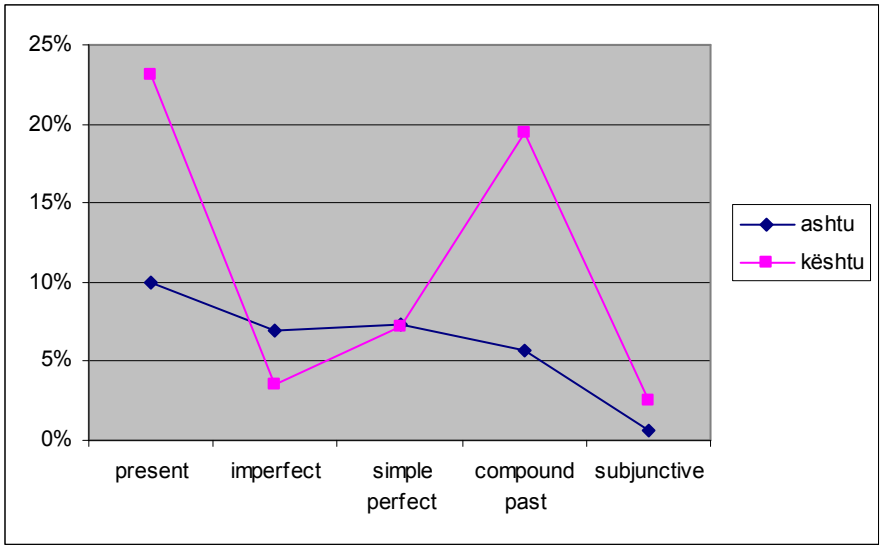


Figure 15: Distribution of tenses between *ashtu* and *kështu*

Missing from the graph above is the large number of nouns (44%) that are collocations of *ashtu* while nouns that collocate with *kështu* are only 17%. According to Biber (1988) this would put *ashtu* in the more informational style and *kështu* in the more affective / involved style.

Deictic adverbs of quantity *aq* and *kaq*

In CALT we count 7081 *aq* 'that much / so' and 4140 *kaq* 'this much / so.' These two words can modify a noun, an adjective, an adverb, or can stand alone in a pronoun like fashion. Çabej (1976a) considers these forms as the oldest of the deictics in *a-* and *k(ë)-*. It has been attested in the first Albanian written document (Buzuku) but the root *q* (or *që* in some dialects) is still seen without the deictic particle in Budi's and Bogdani's later writings. The root form *qyetë* is the reflection of the Indo-European root **quot-* 'how many.'

...ardhje dhe paraqitje	aq	shumë të pritura…
...1,5 $ copën,	aq	sa riprodhohen…
...një realitet artistik, sa interesant	aq	i vërtetë…
...nuk do të merrnin edhe	aq	lehtë kurajë…
...ka po	aq	trajta agresive…
...marrim vendime	kaq	të rëndësishme...
...që merr një post	kaq	të lartë në SHPK…
...Me	kaq	është mbyllur pjesa e parë...
...mendoj se	kaq	mjaftojnë për të nxjerrë përfundimin...
...lexon me	kaq	vëmendje…

Concordance 3: Illustration of the uses of *aq* and *kaq*

The peculiar uses of this adverb are illustrated by the variety of translations one can come up in English[23]. I have tried to translate as conservatively as possible and, still, I could not come up with consistent similar wordings. By looking at the collocations generated by these two words, it can be noticed that they form several prefabs or idiomatic expressions. The most frequent common collocation is the particle *po* as in *po aq/kaq* 'just that/this

[23] …appearances so much anticipated…
…$1.5 per piece, enough to reproduce…
…an artistic reality, as interesting as true…
…they wouldn't be encouraged so easily…
…has so many aggressive traits…
…they make decisions so important…
…that takes a position so high in the SHPK…
…With this much ended the first part…
…I think that this much is enough to conclude…
…reads with so much attention…

much.' Another construct that creates a high scoring collocation is *sa* ___ *aq/kaq* ___ 'as ___ as ___' as in the example …*sa interesant aq i vërtetë*... 'as interesting as true' or *aq/kaq* ___ *sa* (*që*) ___ 'so ___ that ___' as in …*kaq lehtë, sa ata vendosën*... '...so easily, that they decided...'

The most interesting collocations are the accusative prepositions *me* 'with,' *për* 'for' and the nominative *nga* 'from' which form the expressions *me aq/kaq* as in the example *me kaq përfundoi pjesa e parë* which could be translated as 'with that/this ends the first part' or *nga aq/kaq u trembe ti* 'from that/this much you were scared.' Even though it is easy to see a similarity with a pronoun, these constructs are just suspended phrases missing some generic noun like *stuff*, *thing*, etc. The same explanation can be offered for the construct *vetëm kaq* 'only this much' which also behaves like an anaphora as in *Ajo ka pasur dhimbje koke;* ***vetëm kaq*** *dija unë.* 'She had a headache; **only this much/that's all** I knew.'

The distribution of the grammatical categories of the words collocating with *aq* and *kaq* is represented in the table below:

	aq	kaq
present	*16%*	*16%*
past	*18%*	*13%*
noun	*14%*	*17%*
adjective	*34%*	*24%*
other	*18%*	*30%*

Table 34: Grammatical categories for *aq* and *kaq*

There are no big differences except for the larger amount of past tense verbs (dominated by imperfect) and the larger number of adjectives going with *aq*.

By looking at the meaning of the highest collocations there are some interesting differences. The word *aq* has as its top collocated verbs *dukej* 'seemed', *dëshiruar* 'desired', *lejonte* 'allowed.' The word *kaq* instead has *mjaftuar*, *mjafton*, *mjaftohet* 'is enough,' *mbyll* 'close,'

mbaron, *mbaroi* 'ends.' While the collocations for *aq* have meanings of indefiniteness, the collocations for *kaq* point to a closed topology.

Summary

The adverbs of place *këtu*, *aty*, *këtje*, and *atje* are analyzed in more detail as they form the comparative matrix for all our analysis. The discovered patterns correspond to dictionary entries and add to them additional features as well as their significance in naturally running text. Following the traces established by the adverbs of place, the other adverbs – *andej* / *këndej*, *ashtu* / *kështu*, and *aq* / *kaq* – are analyzed and placed in the matrix of the deictic adverbs.

The matrix consists of sliding scales with opposite values. These scales are the following: present tense / past tense; definite / indefinite; person / non person; venitive / andative; familiar / unfamiliar; known / unknown; up / down. The balance moves in varying degrees towards one end or the other but consistently we have the proximal near the left side of the opposition and the distal towards the right side.

While we started our analysis of the adverbs of place as a three degree set, after substantial evidence from data, we concluded that there are indeed two two-term sets, both sets being opposed to each other.

Demonstratives

Diessel (1999:93) distinguishes four pragmatic functions of demonstratives which are exophoric (referring to entities in the speech situation) and endophoric further differentiated in anaphoric (coreferential with a prior NP), discourse deictic (coreferential with a prior proposition), and recognitional which indicates that the hearer can identify the referent based on specific shared knowledge.

Albanian uses the same forms for both determinative and pronominal duties.

Personal and demonstrative pronouns

In this chapter we apply the computational techniques developed in the previous chapters to the Albanian pronominal system. The focus will be in determining the existence of third person personal pronouns and their relationship to distal demonstrative pronouns via quantitative methods. By analyzing collocates and the structures in which these words appear in our 19 million word corpus, we will see that the distributions of what are called third person personal pronouns and demonstrative pronouns are equivalent and discriminating them as separate categories becomes a questionable task.

The reference grammar of the Albanian language (Dhrimo et al. 1986) describes the category of personal pronouns as a set of 1st, 2nd and 3rd person pronouns with their respective definitions of the person that speaks, the person spoken to, and what/who is spoken about. This follows a long tradition started in the second century B.C.E. with Dionysius Thrax' parts of speech in the Art of Grammar (Kemp, A. 1987). 1st and 2nd person pronouns refer to humans and hence the name of the feature "person." Because of its interchangeability with any noun and the distinctions between discourse and story, 3rd person

could best be referred to as non-person (Benveniste, E. 1966) or, as Bhat (2004) prefers, proforms. Between pro-forms though, there still remain deictic features better related to discourse. Even though the contrast between deixis and anaphora has been identified and analyzed since Apollonious Dyscolus' second century C.E. work, there still seems to be confusion in the definitive labeling of these categories. According to Apollonius, anaphora concerns reference to some entity in language, while deixis to some entity outside language (Lehmann, W. 1982). The same categories have been described as endophoric and exophoric references (Halliday & Hasan, 1976). Claude Hagège (1992) includes both of them as the core of a larger and more exhaustive system called anthropophoric. While 1st and 2nd person pronouns are proper deictics or exophoric pronouns, third person suffers from its dual anaphoric and deictic nature making it hard to be classified under one or the other.

The duality of third person – anaphoric and deictic – has become the subject of many studies focusing on one language or across languages. If the pronoun is anaphoric, it is classified as a 3rd person personal pronoun. Languages such as English, Italian or French, complete the paradigm of the personal pronouns with *he*, *lui*, and *il*. If it is deictic, it gets relegated to a whole new set of demonstrative pronouns such as *this/that*, *questo/quello*, and *celui-ci/celui-là*. Grouping separately from demonstratives both anaphoric and third person pronouns is counterintuitive. First, it ignores the anaphoric usage of demonstratives as in the example …*having said all that, what...* where *that* refers to endophoric entities. Second it unifies in the same paradigm 1st and 2nd person pronouns that refer to extra-linguistic actors of the speech act (such as *I* and *you* in English) with intra-linguistic references where the pronoun merely refers to another previously mentioned object as in the overanalyzed donkey sentences *Pedro owns a donkey. He feeds it.* where *he* and *it* refer back to *Pedro* and *donkey*. According to Benveniste (1966:256), la régularité de la structure formelle et une symétrie

d'origine secondaire produisent l'impression de trois personnes coordonnées[24]. While 3rd person personal pronouns are not compatible with the referential terms *I/here/now*, demonstratives that are better related to the speech act are left in a separate paradigm. As always, confusion arises in the middle. From a sample of 225 languages, Bhat (2004) identifies 126 two-person languages with just 1st and 2nd personal pronouns, and 99 three-person languages with a complete set of 1st, 2nd and 3rd person personal pronouns. Languages belonging to two person systems either do not have a third person at all or what is considered as such has close ties to the demonstratives.

Following the above model, Albanian would have a two-person personal pronoun system. However, Albanian reference grammars refer to the deictic usage of pronouns as demonstratives and to their anaphoric usage as 3rd person personal pronouns. Always according to these grammars, the anaphoric usage is limited only to distal demonstratives.

Inventory of personal/demonstrative pronouns in Albanian

The most common forms of the Albanian demonstratives assume both determinative and pronominal functionality.

[24] [translation: AM] The regularity of a formal structure and a symmetry from a secondary origin give the impression of three coordinated persons.

determinative / pronoun	Distals				Proximals			
	Singular		Plural		Singular		Plural	
	M	F	M	F	M	F	M	F
NOM	ai	ajo	ata	ato	ky	kjo	këta	këto
ACC	atë				këtë			
DAT GEN ABL	atij	asaj	atyre		këtij	kësaj	këtyre	

Table 35: Inventory of common Albanian personal/demonstrative pronouns

Personal/demonstrative pronouns in Albanian inflect according to number, gender and case as shown in the table above. In their determinative functionality, their form agrees with the noun they define. As pronouns they take gender, number and case features from the entity they replace or refer to.

While nominative and accusative cases share only their plurals, genitive, dative and ablative share all the forms. The differences between genitive, dative and ablative are syntagmatic: dative is never preceded by a prefix or article and its syntactic role is indirect object; genitive is always preceded by a pre-posed article also known as the particle of concord: *i*, *e*, *të* and *së* and it is usually a possessive construct agreeing with the "possessed" substantive through the particle; ablative is preceded by one of the many prepositions of adverbial origin such as *larg* 'far,' *afër* 'near,' *pranë* 'next to,' *mes/ndërmjet* 'among,' *midis* 'between,' *para* 'before,' *pas* 'after,' *sipas* 'according to,' *prej* 'from, of,' *drejt* 'toward,' *karshi/kundër* 'opposite,' *krahas* 'alongside,' *rreth* 'around,' *brenda* 'inside,' *përveç* 'aside,' *gjatë* 'during,' and *jashtë* 'outside.'

Beside distal and proximal demonstratives beginning, respectively, with *a-* and *k(ë)-*, there is also a second group stripped of the deictic prefixes which can be called non-deictic demonstratives. The reason we keep the category "demonstrative" is that both historically and functionally they preserve the demonstrative feature and can be used exophorically.

While the deictic prefix can be removed transparently for most of the words, it is not so for nominative singular. The remainders of *ai*, *ky*, *ajo*, *kjo* cannot stand by themselves. However, they are found freestanding as preverbal genitive clitics *i* [SG] / *u* [PL] and accusative *e* [SG] / *i* [PL] all derived from the old deicitic-free demonstratives *e-, *i-, *uos-, *uo (Çabej, 1976a:157, 1976b:235). Another pronominal relic of these forms is the usage as particles of concord especially in examples such as *Shtëpia e Agimit është e madhe por e Petritit është më e madhe.* 'The house of Agim is big but [that] of Petrit is bigger.' The sequence *e Petritit*, when analyzed closely (and translated), contains the nominative pronoun *e* followed by the dative *Petritit* which Selman Riza (2002:173) considers as a genitive relict of the old Albanian.

	Singular		Plural	
	M	F	M	F
NOM			ta	to
ACC	të			
DAT				
GEN	tij	saj	tyre	
ABL				

Table 36: "Bounded" pronouns

Even though nominative and accusative demonstratives are part of this group, they can never appear in a sentence as subjects or objects respectively. For this reason the row for dative and nominative singular is grayed out. They can only be found following nominative prepositions *nga* 'from' and *tek/te* 'at, to' or accusative ones *me* 'with,' *mbi* 'on,' *nën* 'under,' *për* 'for,' and *në* 'in.' Since only substantives can take a preposition, these forms can never assume the adjectival role of their corresponding deictic forms. The sentences:

Ai pa* ***ta*. 'he saw **them**'
****To*** *shkuan në Itali*. '**they** went to Italy'

are both ungrammatical because the highlighted forms – the first accusative and the second nominative – cannot be used without the support of a preposition[25] or a deictic particle such as *a-*, *k(ë)-* or *nji-*. The same is true for genitive and ablative cases that are always preceded respectively by an article, like *i*, *e*, *të*, *së* 'of,'[26] or adverbial preposition, *prej* 'from,' *sipas* 'according to,' etc. as in the concordances below.

…Identiteti kombëtar	tek ta	thuajse nuk ekziston…
…për shumë	nga to	shteti është ende në gjyqe…
…në shtëpi kur jetojmë vetëm gra	në të	…
…të gjithë pjesëtarët e familjes	së saj	…
…Disa	prej tyre	u ngjitën edhe në tavan…

Concordance 4: Uses of "bare" demonstratives

Another group of demonstratives that are rarely used and mostly in dialectal or historical documents is formed by ablative forms. By analogy with noun inflection, where the plural indefinite of the ablative is marked by the ending *-sh*, pronouns in this group take the same ending. The existence of this ending constitutes the reason for having a fourth case in Albanian (Friedman 2004).

	determinative singular/plural		pronoun plural	
	M	F	M	F
a-	asi	aso	asish	asosh
k(ë)-	kësi	këso	kësish	kësosh
ø-			sish	sosh
			syresh	

Table 37: Ablative series

[25] In Albanian, nominative can be preceded by prepositions which are *nga* 'from' and *tek* 'at.' Accusative takes *për* 'for', *me* 'with', *në* 'in,' etc.

[26] See Table 3: Some usage examples of the particle of concord in the introductory chapter.

These forms are rare in CALT but searching on Google[27] we found an interesting construct of the "bare" ablative:

…*shumë* ***sish*** *ikin në Austri e në Rusi. Ndërkaq shumë* ***nga ata*** *që mbetën*…
…many **of them** leave for Austria or Russia. Meanwhile many **of them** that were left…

The same meaning is expressed by two different devices, one with the ablative and the other with a nominative preceded by a preposition. The function of this ablative, being a rare construct, is being taken over by the construct preposition + nominative/ablative affirming the move of the Albanian language towards analytic devices described by Likaj (1997).

The inventory of Albanian pronominal forms reveals a tendency to reuse the same forms for both determinative and pronominal functions. The second ablative series shows a specialization of the forms caused by the analogical attachment of the indefinite ablative noun ending *-sh*. Only the pronouns preceded by *a-* or *k(ë)-* can be used in determinative roles. Apparently, *a-* and *k(ë)-* add the adjectival feature to the otherwise pure pronouns.

In Table 35, Table 36 and Table 37, it can be observed that the distribution of gender over number and case is unbalanced. The nominative and ablative have masculine and feminine for both singular and plural. The genitive, dative and ablative have both genders in singular but only one form for plural. In the accusative we observe the opposite distribution with both genders in the plural but only one in the singular, conflicting with Greenberg's universal 45 which says that if there are any gender distinctions in the plural of

---

[27] <http://www.google.com/search?q="shume+sish" &start=10>. January 15, 2007:
Kadare -6 - 9:44pm Shume sish vriten ne kryengritje, shume sish cohen syrgjyne ne Azi, shume sish ikin ne Austri e ne Rusi. Nderkaq shume nga ata qe mbeten ne Malesi, ...
www.cercizloloci.8m.net/photo3.html - 151k - Cached - Similar pages

the pronouns, there are some gender distinctions in the singular also (Greenberg 1966). Plank and Schellinger (1997) found out that there are a considerable number of languages that violate this universal – about 10% of their data set. By including case in their analysis and not just number and person, the Albanian demonstrative pronouns system shows that universal 45 exceptions could be even more.

The origin of Albanian demonstratives

Albanian demonstratives reflect common developments with other Indo-European languages. According to etymological analysis of the personal/demonstrative pronouns in Albanian, their roots are clearly derivations of Indo-European demonstrative roots. Following Çabej (1976:31, 1977:109-110) and B. Demiraj (2002:226), these constructions in Albanian must be quite recent because they have not been subjected to the aphaeresis of the initial unaccented vowel. The common Albanian pattern seen in Latin *amicus* giving Albanian *mik* has not happened in *ata*, *ato*, *atij*, *asaj*, and *atyre*. This conclusion is disputed by the proposals of Orel (1998) and Hamp (p.c.) that propose that the older deictic particles *a-* and *kë-* existed independently of the pronominal system and, through frequent usage in a preposed position merged later into one word (see discussion of Table 32 in the adverbs section).

By observing the two parallel paradigms, distal and proximal in Table 35, *a-* and *k(ë)-* can be identified as prefixes attached to the pronominal roots. The pronominal roots, or what is represented in Table 36 as non-deictic, are found unbound, without the prefixes *a-* or *kë-*, in 16th century writings. Today, these roots, *tij*, *saj*, *tyre*, *të*, *ta*, and *to*, can be found unbound only when they are preceded by a preposition or article. This would mean that

instead of the prefix, they are "bound" to a preposition or pre-posed article. The old ablatives *sish*, *sosh*, *syresh* are an exception.

There are not many studies dealing with the etymology of the pronominal part of the demonstrative and even less are concerned with the deictic prefixes. Çabej sees the prefixes *a-* and *kë-* as hypercharacterization devices inferring that the pronominal part already had a demonstrative functionality. This hypercharacterization, apparently in analogy with the deictic adverbs of place, added granularity to an already existing system. Furthermore, *njito* 'these (feminine)' or *njita* 'these (masculine)' show how loosely attached the deictic prefixes are. The prefixes *a-* and *k(ë)-* are easily replaced when the deictic particle *nji*, equivalent of *ecco* in Italian or вот in Russian, is attached in front of the pronoun. The particle *nji* has nothing to do with distance therefore reducing *ata/ato* 'those masculine/those feminine' and *këta/këto* 'these masculine/these feminine' to degree-less demonstratives. Çabej concludes that it is not the prefixes that transform them into demonstratives – they were demonstratives all along.

Historical linguist Shaban Demiraj (2002), analyzing the pronominal clitics in Albanian, concludes that they do derive from some disappeared set of personal pronouns. As for the demonstratives, he thinks that their different forms derive from a mix of different Indo-European demonstrative sets but that these words still do not have a clear origin. Bokshi (2004) instead concludes that there has been a unidirectional movement from demonstratives to personal pronouns. The first series of demonstratives deriving from the Indo-European demonstratives, with time, lost its deicticity and constituted the personal pronoun series. The two deictic prefixes were needed to reconstitute the demonstrative

pronouns from these personal pronouns. Following the same pattern, he sees today a new move of distal demonstratives towards third person personal pronouns.

The conclusion that can be reached by these analyses is that old Indo-European demonstratives retained their demonstrative traits in Albanian and, in addition, reinforced their deicticity with the more visible deictic prefixes. As the language evolved, there has been a movement from personal pronouns to clitics and from demonstratives to personal pronouns. The deictic prefixes, *a-* for distals and *k(ë)-* for proximals, are attached not only to old demonstratives but to other pronouns and adverbs as well: *atillë/këtillë* 'such as that/such as this,' *këtu / aty* , *këtje / atje* 'here / there,' *andej/këndej* 'from there/from here,' *aq/kaq* 'that much/this much' and *ashtu/kështu* 'that way/this way.' In *akëcili/akëkush* 'whoever' both prefixes are attached to achieve indefiniteness. On this aspect, the Albanian deictic system has very similar constructs to the Hittite deictics that precede the same category of words with either *ap-* for distals or *k-* for proximals (Sturtevant 1951:102-115 and Benveniste 1962:66-77).

Third person personal pronouns

From the synchronic point of view, by labeling the distal demonstratives (those that start in *a-*) as personal pronouns, Albanian grammarians needed to establish a set of rules for distinguishing them from each-other. The reference grammar of Albanian (Dhrimo, A. et al. 1986) provides two tests to achieve this distinction.

According to the reference grammar, these pronouns should be called personal when they replace a noun mentioned earlier, giving it a clear anaphoric function. But a quick corpus search will show that Albanian uses pronouns with both prefixes (*a-* and *k(ë)-*) in

anaphoric functions. Furthermore, when needed to resolve antecedent ambiguity in text, Albanian does use the deictic features, as in "the former/the latter" in English. This logic could lead to the conclusion that the personal pronoun paradigm is in fact richer and contains both *a-* and *kë-* pronouns (Murzaku 1989).

...Koshtunica nuk preku ... për më tepër	*ai*	*u përpoq ta mënjanojë Gjingjiçin, por*	*ky*	*arriti...*
	the former that one		the latter this one	
...Kostunica didn't touch ... furthermore	he	tried to put aside Djindjic, but	he	achieved...

It is obvious that the second pronoun, having multiple possible antecedents, needs some other tool to differentiate it. By using the proximal demonstrative in opposition to the distal demonstrative, anaphora ambiguity is resolved with the calculation of distance inside the text.

The other test suggested by the grammar is that the use of the pronoun without the leading *a-* is an indicator that we have a personal pronoun rather than a demonstrative. This test seems to suggest that, if the non-deictic root of the pronoun is a personal pronoun, then, anything it replaces is also a personal pronoun. Submitting a phrase search to any search engine, it can be seen that not only pronouns starting in *a-* can fill this slot. This search retrieved 5300 "*me ta*", 3000 "*me ata*" and 500 "*me këta*" in very similar syntactic structures.

...në suaza të Komisionit dhe i cili punon	*me ta*	*çdo ditë…*
...in subgroups of the Commission and which works	with them	every day…
…se ky më shumë rri ... e punon	*me ata*	…
	with them	
…because he/this one mostly stays … and works	those ones	…
…më pas dërgon një koreograf, i cili punon	*me këta*	…
	with them	
...later sent a choreographer, who works	these ones	…

In the examples above, 'with them' is part of identical structures differing only in the use of the pronoun *me ta* a non-deictic, *me ata* a distal and *me këta* a proximal. It is obvious in this case that both, distal and proximal demonstratives, can be replaced by the corresponding non-deictic pronoun.

Both tests of pronoun status suggested by the reference grammar of Albanian, their anaphoric role and their substitutability, are rather ineffective in discerning personal from demonstrative pronouns.

Quantitative analysis

Neither diachronic nor synchronic analyses until now have provided a good answer to our original question of whether there is a 3rd person personal pronoun in Albanian. Etymologically, there seems to be a constant move between these demonstrative and personal pronouns without a definitive answer on the origin of the deictic prefixes *a-* and *k(ë)-*. On the other hand, today's descriptive studies offer no clear division between personal and demonstrative pronouns. A part of speech is defined by the meaning and by the role that a word or a group of words play in a sentence. While the introspective and diachronic analyses can provide good explanations and descriptions of the meaning as well as functionality and origin of these words, a quantitative analysis could complete it with a better view of how these forms are distributed in today's usage and what patterns they create in natural text. Following Firth's (1957) slogan "you shall know a word by the company it keeps," this new dimension, based on large scale data, brings additional arguments to the suggestion that today's Albanian is indeed a two person language and that the line of demarcation being sought between personal and demonstratives pronouns perhaps does not exist.

Analyzing the semantic content of the pronouns in question, the working hypothesis is that distal and proximal demonstratives are associated with words belonging to their respective deictic dimensions.

Pronouns, which are the object of this study, are function words and the quality of collocations for such words should not be affected by the situation of the lexicon in general and the language registers so, we are confident that the acquired results can be trusted.

Discussion of results

The project aimed at two separate results. The first one was to create tools and datasets that would provide clean concordances and statistical data for our study. About 324,000 concordance lines (160 characters each) and the frequencies in the following table were generated for the eight *a-* pronouns and the corresponding *k(ë)-* pronouns of today's Albanian.

The table below includes absolute numbers as they were found in CALT. Even at this stage, these numbers have an inherent value that needs explaining.

Distal			Proximal		
	occurrences	collocations		occurrences	collocations
ai	43,310	5,492	ky	17,502	2,456
ajo	19,890	2,812	kjo	26,399	3,347
atë	11,228	3,269	këtë	62,972	7,245
ata	22,803	3,178	këta	4,610	638
ato	15,865	2,296	këto	21,508	2,909
atij	4,219	601	këtij	22,809	3,195
asaj	4,604	657	kësaj	20,048	2,886
atyre	5,319	775	këtyre	9,686	1,464
total	*139,210*	*19,080*	*total*	*185,554*	*24,140*

Table 38: Absolute frequencies of the *a-* and *kë-* pronouns

If distal forms were personal pronouns as well as demonstratives, this double duty would have implied that their frequency should have been higher. At a first glance, proximal demonstratives have almost 30% more occurrences than the distal demonstratives (185,554 proximals vs. 139,210 distals). By analyzing the data in more detail, we see that the distribution among the several forms is uneven with respect to case, gender and number. The pronoun *ai* (nominative, singular, masculine, distal) occurs twice as much compared to *ky* (nominative, singular, masculine, proximal). But the corresponding feminine forms *ajo* and *kjo* are more evenly distributed with *kjo* occurring only 30% more than the distal form. The same distribution can be seen for their corresponding feminine plural forms *ato* and *këto*. The distribution of masculine plural forms *ata* and *këta* is inversed with the distal form having five times more occurrences than the proximal. It should be noted that these forms are shared between nominative and accusative. Singular accusative and genitive/dative/ablative both singular and plural are heavily unbalanced in favor of the proximal forms.

Through further analysis of the concordances, it is observed that proximal and distal demonstratives have inverted ratios in their uses as pronouns and determinatives. As illustrated in the table below, *a-* words appear mostly in pronoun roles while *k(ë)-* words in determinative ones.

	pronoun	determinative
ai	**98**	2
ky	29	**71**
ajo	**92**	8
kjo	48	**52**
atij	**68**	32
këtij	4	**96**
asaj	**60**	40
kësaj	8	**92**
atë	**71**	29
këtë	14	**86**
ata	**96**	4

këta	28	**72**
ato	**73**	27
këto	12	**88**
atyre	**78**	22
këtyre	2	**98**

Table 39: Pronoun to determinative ratios

These ratios are also illustrated by a random selection of *ky* and *ai* concordances from CALT:

11167	…të gjithë shkencëtarët dhe teknikët, dhe	ky	është një kusht i domosdoshëm…
11168	…mban gjalle optimizmin se edhe	ky	moszhvillim i deritanishem i turizmit…
11169	Më në fund,	ky	libër nuk është shkruar vetëm për ata…
11170	…një diletanti politik, qoftë	ky	edhe një personazh "legjendar e i stërnjohur"…
11171	…sukses ideali i ekumenizmit, atëherë	ky	vend, kjo shoqëri është Shqipëria…
11172	…shfaqet në rolin e Konstatinit, ndonëse	ky	i fundit prirej nga parimet pragmatike…

Concordance 5: Random selection for *ky*[28]

4272	…të ndërtohej një "qiellgërvishtës",	ai	që më vonë do ta njihnim si "15-katëshi".
4273	…për t'iu falur Zotit,	ai	sens pra do t'u jepte organizatorëve…
4274	Deri në datën 30 maj	ai	iu ka siguruar 5 vendparkime para shkollës…
4275	…njerëzit e tij e bëjnë që, sidoqoftë,	ai	ta ketë mirë me kryeministrat…
4276	…që thotë se	ai	ngjiste si hale me një copë mëndafshi.
4277	Bombë Ylli Pangos,	ai	profesor akademik kishte një…

Concordance 6: Random selection for *ai*[29]

In the concordances above, even though the functionality seems the same, the ratios are completely opposed. As discussed in the chapter Discussion of the Method, referring to Jurafski et al. (2001), Gahl and Garsney (2004), and Bybee (2006), frequency of usage does influence meaning. The high frequency of the *k(ë)-* words in determinative roles, almost

[28] 11167: …he (this one) is…
11168: …this non-development…
11169: …this book…
11170: …this being even a character…
11171: …this country…
11172: … he/this last one - the latter…

[29] 4272: …it (that one) which later…
4273: …that meaning…
4274: …he (that one) made available to them…
4275: …he (that one) should have a good relationship with…
4276: …he (that one) looked like…
4277: …that academic professor…

transforms them in markers for definiteness. When the *a-* words are used as determinatives, their rarity puts in evidence the distal role rather than the role of the determinant.

The next step in the data analysis is to move the focus out of the words themselves and observe the associations that emerge between the target words and words in their neighborhood. Through the use of the mathematical apparatus described earlier, these lexical-grammatical associations help define their similarities or differences. Word neighborhood (or span) is defined as the number of words on each side. By using a right and left span of four words and looking for collocations only with words that have frequencies greater than five, substantial lists of collocates for each of the pronouns were generated. Each target or nucleus word generated between six hundred and eight thousand collocates in proportion to the type and frequency of the word.

Once the data was acquired, it was expected that some results would correspond to the initial hypothesis, i.e. that distal and proximal demonstratives are associated with words belonging to their respective deictic dimensions.

The demonstratives ai *and* ky

The patterns that were observed in the study of the adverbs *këtu*, *aty*, and *atje* are also noticeable in the data extracted from the collocations analysis for *ai* and *ky*.

The distribution of grammatical tenses[30] in the table below follows a similar distribution to the tenses of the verbs collocated with the locative adverbs.

tenses	ai		ky	
	occurrences	percentage	occurrences	percentage
present	97	10%	128	26%
subjunctive/future	51	5%	21	4%
simple perfect	121	11%	36	4%
compound past	110	12%	21	7%
imperfect	160	16%	20	4%

Table 40: Distribution of verbal tenses among *ai* and *ky*

The numbers follow a diverging pattern where *ai* has the highest number of collocations with verbs in the past tense and especially with imperfect and *ky* is found mostly close to verbs in the present tense. This divergence is better visible in the graph below:

[30] Since the collocational analysis is based on single words, the most precise information that can be gathered is for tenses that are formed through synthetic devices such as present, simple past and imperfect. Compound past tenses are formed analytically by an auxiliary and a participle. Their numbers are deduced from the number of identifiable past participles. The several kinds of compound pasts cannot be set apart with this kind of approach. Future tense is also formed analytically by preceding conjunctive with *do*. Conjunctive has a particular verbal form preceded by the particle *të*. It is this particular form that is counted in our collocations.

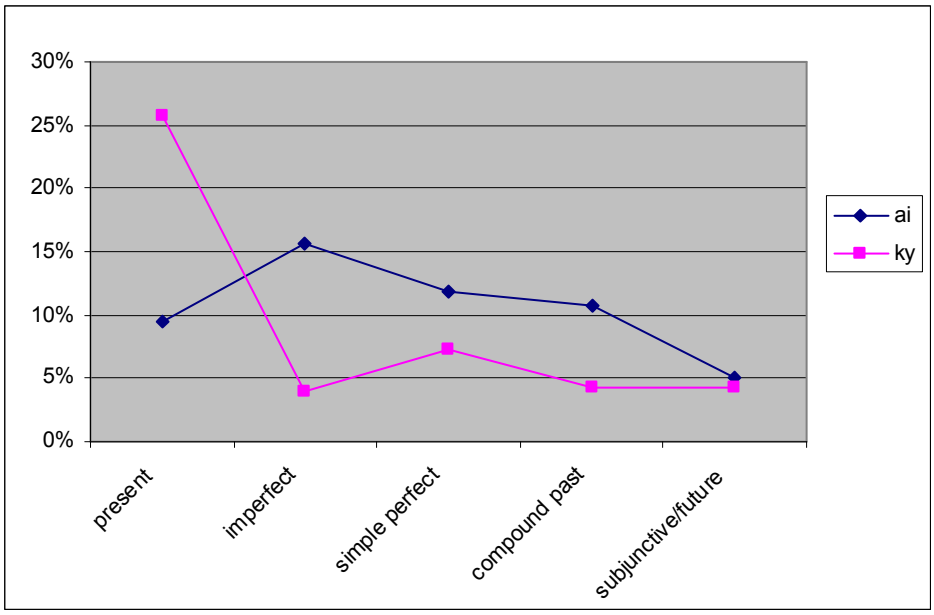


Figure 16: Diverging lines of verbal tenses for *ai* and *ky*

The graph in Figure 16 highlights the similarity of distribution of tenses with those in Figure 12 (adverbs of place) where *ky* matches with *këtu* and *ai* with *atje*. The conclusion is that the proximal *ky* is closely related to the speech act or to the deictic center and *ai* to storytelling.

Combining these data with the grammatical category (pronoun vs. determinative), the reference to the moment of speech is further reinforced.

	pronouns	determinatives
ai	42,444	866
ky	5,076	12,426

Table 41: Pronominal and determinative usage of *ai* and *ky*

The demonstrative *ky*, by sheer force of numbers, takes a role similar to the determinative article: something that is known, visible, in front of the dialogue participants is determined by the determinative use of the proximal. In the small number of cases where *ky* is used as a pronoun, it becomes a pure proximal whether it is referring to an antecedent that is closer to the pronoun (as in English 'the latter') or an object in the extratextual world that

is closer to the *origo*. Objects that need to be put in evidence for the listener are referred to by the pronoun *ky*.

The demonstrative *ai*, through the overwhelming usage as a pronoun, takes a more significant role in fulfilling anaphoric functions. In the few cases it is used as a determinative, it marks distance. The noun that is determined by the distal *ai*, refers usually to an entity far from the *origo* or by extension not well known or not visible. The following example illustrates the meanings of *ai* and *ky* very well:

…*ai*	*nuk ishte*	*ky.*	*Ai*	*ishte*	*ai*	*tjetri.*
…that one	was not	this one.	that one	was	that	other (one).
…he		he.	he		he	

'…he was not this one. He was the other one.'

In the first sentence *ai* is simply an anaphoric while *ky* refers to someone in front and/or visible from the speaker's point of view (= close to *origo*.) In the second sentence, the first *ai* remains an anaphoric while the second one's relationship to the *origo* is simply a non-relation. By itself, *tjetri* already means 'the other one.' Being preceded by ai adds to the indeterminateness of the overall reference. The following table of collocations shows how *ai* and *ky* have different collocation scores with words from the 'other' family:

other[31]		ai			ky		
singular	tjetër (adj)	3.28	3.86	4.25	3.09	0.81	0.84
	tjetra	2.55	-0.97	-0.83			
	tjetrën	2.84	-0.43	-0.41			
	tjetri	3.24	0.71	0.77	3.62	1.15	1.43
	tjetrin	2.52	-1.99	-1.69			
	tjetrit	2.71	-0.94	-0.85	2.43	-1.18	-0.97
plural	tjera (adj)	2.83	-2.25	-2.12	2.31	-5.73	-4.51
	tjerash	3.53	1.07	1.28	4.25	1.64	2.54
	tjerave	5.08	3.89	7.99			
	tjerë (adj)	2.85	-1.40	-1.33	2.22	-4.77	-3.64
	tjerët	3.58	2.66	3.25	2.35	-1.89	-1.51
	tjerëve	3.46	1.32	1.54			

Table 42: Collocations of 'other' with *ai* and *ky*

The collocation scores in Table 42 are not spectacular but what attracts attention is the fact that *ky* has either very low scores or no collocations at all with the various forms of 'other.'

Continuing our analysis we drill down to look for similar details in the following two tables which include the top thirty collocates for *ky* and *ai*.

Collocate	English	MI-score	T-score	Z-score
fundit	(the) last	6.00	31.97	90.60
yni	(the) ours	7.50	4.87	23.17
zbatohet	(is) implemented	6.32	5.55	17.51
zgjidhet	(is) elected	5.93	6.44	17.78
pikërisht	exactly	5.47	12.96	30.49

[31] The word for 'other' in Albanian has a peculiar behavior. The singular adjectival form is *tjetër* which remains the same for both masculine and feminine, e.g. *djali tjetër* / *vajza tjetër* 'the other boy / the other girl.' In the plural, it takes respectively the masculine and feminine forms *tjerë* and *tjera*. These forms though cannot stand without the support of the particles of concord derived from the old demonstratives as in: *djemtë e tjerë* / *vajzat e tjera* (NOM), *djemve të tjerë* / *vajzave të tjera* (GEN/DAT/ABL). Both singular and plural forms can be nominalized by being attached nominal declension endings and being used as pronouns. The plural forms maintain their particles of concord even when they are nominalized, e.g. *të tjerët*, *së tjerash*, *të tjerave*, etc.

rasti	(the) case	5.69	7.46	18.94
miratohet	(is) approved	6.02	5.04	14.36
shtrihet	(is) extended	5.89	4.15	11.31
fillimi	(the) beginning	5.46	5.24	12.28
thelbi	(the) essence	6.12	3.76	11.09
thelbësor	essential	6.56	3.42	11.76
realiteti	(the) reality	5.78	4.10	10.73
vendimtar	decisive	6.43	3.39	11.15
mendoj	(I) think	5.08	8.50	17.48
shkaku	(the) cause	5.37	4.91	11.16
realizohet	(is) realized	5.39	4.79	10.97
mendoni	(you) think	5.33	5.01	11.24
kundërshtua	opposed	7.24	2.99	13.00
përbën	makes up	5.24	5.34	11.61
shërbejë	serves (Subjunctive)	5.54	4.05	9.77
problem	problem	5.02	7.60	15.30
përse	why	5.02	7.53	15.14
arrihet	(is) achieved	5.39	4.50	10.30
vlerësim	valuation	5.63	3.84	9.58
qenka	is (Evidential)	5.84	3.55	9.47
miratua	approved	5.90	3.46	9.46
kadriun	proper noun	7.23	2.84	12.31
proces	process	4.99	6.60	13.15
ekziston	exists	5.08	5.71	11.75
gabim	mistake	5.34	4.32	9.72

Table 43: Top thirty collocates for *ky*

The most evident features of the collocations of the demonstrative *ky* in Table 43 are: 1) present tense verbs, 2) definite nouns (8-to-3 against indefinite nouns), 3) determinative expressions (such as *i fundit* 'the last' and *yni* 'ours'), and 4) evidential/admirative mood[32] for the verb 'to be' *qenka*.

[32] Friedman (1981) argues the category marked by evidential/admirative is status, which is not itself modal, although it can interact with mood. He supports this argument with the facts that admirative takes the indicative negator *nuk* rather than the modal negator *mos*, and, moreover, that the admirative can occur in modal as well as indicative constructions, e.g., in analytic subjunctive constructions in *të* (*nuk mund të dilkemi këndej* 'we can't get out this way' [when arriving in front of a locked door that was supposed to be open]). If the admirative were itself a mood, it would require a double layer of modality to account for such usage (Friedman 2001).

As discussed in a previous article (Murzaku 1989), the coexistence of the determinative *ky* with a definite noun is explained by the extralinguistic information that is conveyed by the deixis. An object determined by *ky*, probably, is also visible to the participants in the dialogue or is determined by the context in discourse. Çabej (1976:31, 1977:109-110) considers the deictic prefixes as hypercharacterization devices that simply add granularity to the determinative functions of the demonstratives or other deictic adverbs.

The higher ratio of nouns in the top thirty collocations (37%) vs. the overall ratio among *ky*'s collocations (28%) supports the adjectival role of *ky*. Among the top thirty collocations of *ai*, the ratio of nouns is much lower than the rest of *ai*'s collocations (13% in the top thirty vs. 31% overall). This exclusivity relation between nouns and pronouns suggests a predominant anaphoric role for *ai*.

Collocate	English	MI-score	T-score	Z-score
shtoi	added	6.43	16.21	53.25
quajti	called	6.54	10.26	34.95
quan	calls	6.23	8.52	26.13
tha	said	5.67	44.65	112.76
përgjigj	answered	6.03	8.69	24.82
tej	over there	5.59	19.69	48.40
shton	adds	5.76	10.58	27.58
shpreh	expresses	5.40	17.76	40.84
largohej	leaves (imperfect)	5.92	7.05	19.40
intervistë	interview	5.46	9.99	23.41
shtuar	added	5.36	12.13	27.50
largua	left	5.69	6.97	17.70
cilësoi	qualified	5.97	6.04	16.91
dinte	knows (imperfect)	5.45	8.88	20.76
vijoi	continued	5.95	5.64	15.72
kuptonte	understand (imperfect)	6.09	5.07	14.82
vazhdoi	continued	5.34	6.76	15.20
pyeti	asked	5.35	6.38	14.43
thotë	says	4.90	23.78	45.88
keqtrajtonte	maltreats (imperfect)	6.89	3.73	14.39
përgjigjej	answers (imperfect)	5.78	4.35	11.40

thoshte	says (imperfect)	5.00	8.72	17.46
institucionit	institution (DAT)	5.25	5.96	12.99
panoramën	panorama (ACC)	5.74	4.34	11.22
shihte	sees (imperfect)	5.38	5.05	11.53
madje	furthermore	4.83	16.30	30.77
arrestohej	(is) arrested (imperfect)	6.71	3.58	12.92
faktin	fact (ACC)	4.90	11.21	21.64
kthye	returned	5.09	6.26	12.90
linte	leaves (imperfect)	5.24	5.35	11.62

Table 44: Top thirty collocates for *ai*

From the collocations of *ai* in Table 44 emerges the prominence of verbs (24/30) in past (20/24) imperfect (9/20) tenses. The few present tense verbs are specialized in conveying speech acts (*thotë* 'says,' *shton* 'adds,' *shpreh* 'expresses').

Unlike the collocations of *ky*, all verbs collocated with *ai* are in the third person singular implying that *ai* is their subject. One of the initial hypotheses was that pronouns from both paradigms can be found in the same functional slots. The verb *është* 'is' has the same very high collocation values with *ai* (MI:3.23, T:8.64, Z:9.41) and *ky* (MI:4.33, T:32.30, Z:51.41). Other verbs such as *ka* 'has' have similarly high correlations (MI:3.41, T:16.76, Z:19.44 for *ai* and MI:3.77, T:19.92, Z:26.12 for *ky*) thus implying that, at least in the subject role, *ai* and *ky* should be equally distributed.

The difference in verbal form distribution is the existence of several first or second person verbs among the top collocations of *ky* such as *mendoj* 'I think' and *mendoni* 'you think.' The meaning of these verbs is to convey something subjective, – the opinion of the speaker for example and, therefore, implying proximity to the *origo*. The first and the second person of the verbs contribute by themselves to the proximity with the *origo* too. The differences between the two tables show that there is a differentiation in the meaning and functionality of the proximal and the distal and that this differentiation is based on the role of the *origo*.

While *ky* is used mainly as a determinative, *ai* is the anaphoric pronoun of choice. When *ky* is used as an anaphoric, its main role is deictic (inside or outside the text). When *ai* is used as an adjective (determinant), the predominant feature is that of distance. The conclusion that can be proposed here is that both words have deixis as a strong semantic component. High frequency of usage in one or the other role gives prominence to the syntactic functionality thus moving towards grammaticalization of that function and reduction of their respective semantic features. This phenomenon must have already taken place in Albanian once. Indeed, the postposed determinative article has a clear demonstrative origin. Also, the absence of third person personal pronouns in Albanian (Murzaku 2007) underscores the need for borrowing from the demonstratives' paradigm.

Meanwhile, the differences that emerged in the analysis of the locative adverbs are outlined again in the collocations of *ai* and *ky* in which the most important ones are those with verbal tenses: On one hand, both *ky* and *këtu* collocate with the present tense. On the other hand, both *ai* and *atje* collocate with imperfect (and other past tenses).

Continuing with the same kind of analysis, the following tables illustrate other semantic facts extracted from the collocations of *ai* and *ky*.

	AI			KY		
	MI-score	T-score	Z-score	MI-score	T-score	Z-score
andej 'that way'	3.66	1.91	2.40			
asaj 'her DAT'	3.27	1.73	1.89	2.21	-3.25	-2.48
ashtu 'in that manner'	3.77	7.25	9.45	2.55	-2.68	-2.30
atë 'her/him ACC'	3.57	8.36	10.21	2.19	-7.58	-5.72
atje 'there'	3.93	6.44	8.89	2.36	-2.81	-2.25
aty 'there'	3.65	5.25	6.58	2.47	-2.71	-2.25
diçka 'something'	3.45	2.59	3.03	2.35	-2.40	-1.92
dikë 'someone'	4.09	1.50	2.19			
diku 'somewhere'	2.90	-0.31	-0.30			
dikur 'sometime'	3.56	1.80	2.19	2.92	-0.17	-0.16
dikurshëm 'of the past'	3.85	1.09	1.46			
disa 'some'	3.01	0.12	0.12	2.37	-6.55	-5.26
dje 'yesterday'	2.92	-1.11	-1.08	2.25	-6.79	-5.23
tej 'beyond'	5.59	19.69	48.40	3.51	2.16	2.58
tutje 'far away'	3.56	1.40	1.71			

Table 45: Collocation table for *ai* and *ky* with words referring to indefinite or distant objects.

One of the initial hypothesis was that the proximal demonstrative *ky* 'he/this (one)' should have high collocation values with words distributed close to the *origo* I/HERE/NOW and the distal *ai* 'he/that (one)' with words far from the center of the speaking act such as THERE/THEN. In Table 45 *ai* realizes indeed much higher scores than *ky* with words that have distance from the *origo* as one of their semantic features. The distal (*ai*) does have higher collocation values with *atje* 'there,' *tej* 'beyond' and *dje* 'yesterday' as well as exclusive collocation with *tutje* 'far away.'

	AI			KY		
	MI-score	T-score	Z-score	MI-score	T-score	Z-score
imi 'of mine'				6.02	2.63	7.48
kaq 'this much'	3.28	1.74	1.92	4.18	4.85	7.31
kështu 'in this manner'	3.53	6.07	7.29	3.31	2.24	2.49
këtë 'him/this.ACC'	3.71	17.31	22.19	2.42	-8.99	-7.36
këtij 'him/this.DAT'	2.53	-6.63	-5.64	2.79	-1.91	-1.78
këtu 'here'	3.31	2.64	2.94	2.72	-1.55	-1.40
kjo 'she/this.NOM'	2.42	-11.22	-9.17	2.66	-4.11	-3.66
sot 'today'	2.95	-0.48	-0.47	3.31	2.08	2.31
sotmi 'of today'				9.19	2.61	22.30
tani 'now'	3.50	5.36	6.39	3.29	1.96	2.17
tanimë 'already'	2.92	-0.15	-0.14	4.59	2.00	3.48
tashmë 'already'	3.32	2.80	3.12	3.89	5.08	6.91
unë 'I'	3.44	5.96	6.93	3.20	1.75	1.88
ynë 'our'	3.39	1.79	2.04	4.29	3.88	6.06
yni 'of ours'	4.30	1.57	2.46	7.50	4.87	23.17

Table 46: Collocation table for *ai* and *ky* with words referring to definite or close objects.

The demonstrative *ky* does have high collocation values with *unë* 'I,' *ynë* 'our,' *tani* 'now,' *sot* 'today,' *tashmë* 'already.' It also has exclusive collocations with *imi* and *sotmi* which are part of the nominalized adjectives *i imi* 'my-> mine,' *i sotmi* 'today-> today's one.' Both these expressions serve as determiners for *ky* as in *ky i imi* 'this one of mine' *ky i sotmi* 'this one of today.'

Some of the collocations for the proximal have similar scores with the distal. One of the possible causes is the much higher frequency of *ai* as a pronoun and their usage in a larger variety of situations.

An exception to the expectation was the collocation of *ai* with *këtë* 'him/this one' and *këtu* 'here' which have much higher scores than those obtained by *ky*. From a look at the concordances, a plausible explanation can be found based on the high frequency of narrative structures like:

…Shkodra pati fatin të ketë një artist të përmasave të tilla… Pikërisht **këtu** **ai** mësoi edhe ABC-në e parë në pikturë… '…Shkodra was lucky to have an actor of such caliber… Right **here** **he/that one** learned his first ABC in painting…'
… do të vijë një ditë që të tërhiqet nga këto vendime. **Këtë** **ai** e vërteton me faktin… '…one day will come that he will regret these decisions. **He/that one** verifies **this** with the fact…'

In the first type of sentences (…*këtu ai*…), the writer refers to the place where he (the writer) is writing (*origo*). The second type (…*këtë ai*…), as discussed in more length in Murzaku (1990), is a quite common endophoric deictic reference. *Këtë* refers to the latest text unit preceding the demonstrative and is always feminine referring to the complete phrase *këtë gjë* 'this thing.'

While neither of these structures contradicts the collocational analysis up to now, the support for the initial hypothesis by single words is not very strong. Instead, the grammatical features such as tense or abstract semantic features such as 'other' in Table 42 are the strongest collocations contributing to the meaning of these words.

The demonstratives ajo *and* kjo

The demonstratives *ajo* and *kjo* have the attributes of feminine, singular, and nominative. Just like *ai*, *ajo* too is predominantly a pronoun. The proximal *kjo*, unlike *ky*, is evenly split between its usage as a pronoun and as a determinative.

	pronoun	determinative
ajo	18,299	1,591
kjo	12,672	13,727

Table 47: Pronominal and determinative usage of *ajo* and *kjo*

The gravitational pattern that has emerged between *k(ë)-* words and present tense continues in the collocations of *kjo* as well. As the graph below illustrates, *kjo* and *ajo* generate a crossing pattern as in *ai/ky* and *këtu/aty/atje*:

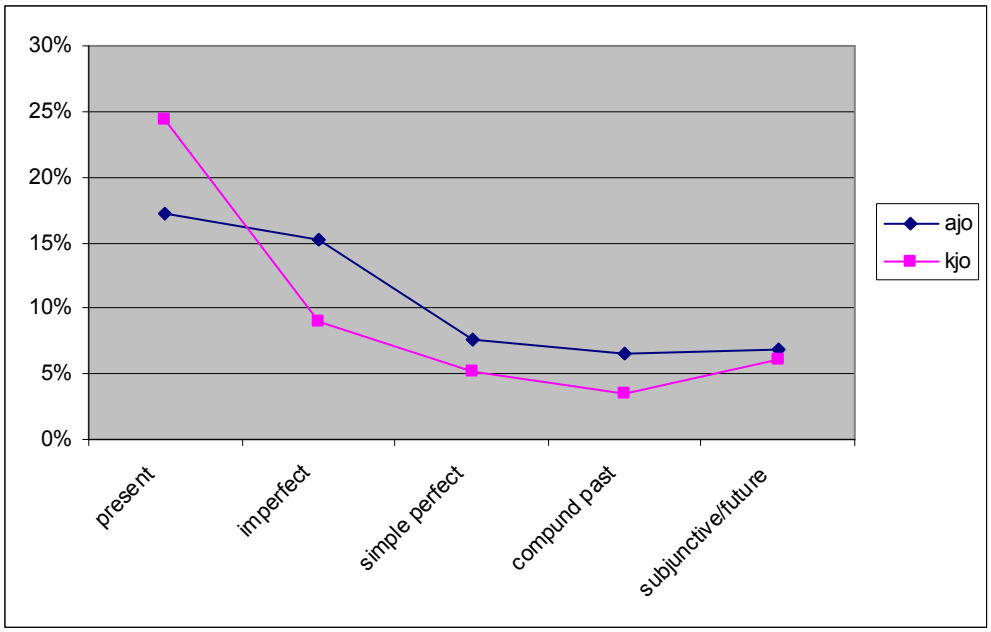


Figure 17: Distribution of tenses between *ajo* and *kjo*

Even though the crossing pattern in Figure 17 is similar in shape to Figure 16, the divergence is smoother for *ajo*. This coincides with the similar numbers in absolute occurrences between *ajo* and *kjo*. While *ai* and *ky* have a ratio of 3-to-1, *ajo* and *kjo* have a 3-to-4 ratio (18,980 vs. 26,399).

As for semantic differences between the distal and proximal forms of the demonstratives, the following tables illustrate and reinforce the patterns that have emerged.

Collocate	English	MI-score	T-score	Z-score
kryesorja	(the) main one	7.31	5.29	23.59
ndodhi	happened	5.68	11.63	29.46
mendoi	thought	6.55	5.56	19.01
shqetëson	disturbs	6.97	4.86	19.24
quan	calls	6.13	5.53	16.38
djeshmja	(the) yesterday's one	9.20	3.95	33.82
shton	adds	5.53	6.40	15.37
rëndësishmja	(the) important one	6.85	3.84	14.56
çani	proper noun	8.06	3.36	19.39
përshtypje	impression	5.58	4.85	11.85
diskutueshme	arguable	6.81	3.48	13.02
ndryshonte	modify (imperfect)	6.25	3.69	11.36
pyeti	asked	5.50	4.66	11.07

pritej	expect (imperfect)	5.59	4.34	10.65
ndodhte	happen (imperfect)	5.27	5.20	11.43
quhet	is called	5.15	5.85	12.34
fëmijën	(the) child (ACC)	5.71	3.97	10.17
thotë	says	4.87	15.92	30.49
shoqin	(the) husband/spouse (ACC)	6.41	3.39	11.06
pëshpëriti	whispered	7.80	3.05	16.06
thoshte	says (imperfect)	5.07	6.14	12.60
vetvete	self	5.55	4.06	9.82
mungonte	misses (imperfect)	5.64	3.66	9.13
përjetonte	relived (imperfect)	7.31	2.85	12.71
armëmbajtjes	the weapon carrying (DAT)	6.50	3.02	10.17
prapë	again	5.14	4.58	9.61
leri	you leave	8.20	2.75	16.68
ndiente	feels (imperfect)	5.86	3.34	8.99
burrin	(the) man/husband (ACC)	5.61	3.55	8.78
mungon	misses	5.09	4.65	9.61

Table 48: Top thirty collocations for *ajo*

Besides the domination of past tenses and third person singular, Table 48 contains no words that would support subjective points of view or closeness to the *origo*.

Collocate	English	MI-score	T-score	Z-score
ndodhë	happens (subjunctive)	6.98	12.46	49.53
arsyeja	(the) reason (NOM)	6.24	9.47	29.14
ndodhte	happens (IMP)	6.19	9.26	27.99
mjaftonte	is enough (IMP)	7.16	6.26	26.44
ndodh	happens	5.83	11.24	30.02
arrihet	(is) achieved	6.18	8.01	24.10
pikërisht	precisely	5.56	16.62	40.34
jona	(the) ours	6.93	5.53	21.55
realizohej	is realized (IMP)	7.00	5.22	20.90
sotmja	(the) today's one	8.12	4.56	26.84
vërtetë	true	5.39	16.03	36.78
realizohet	(is) realized	5.80	7.16	18.90
zgjidhet	(is) resolved	5.74	7.27	18.82
ndikojë	influence (subjunctive)	6.20	5.78	17.54
mirëpo	but	5.56	8.10	19.68
shifër	digit	6.08	6.05	17.60
mendoj	I think	5.36	12.13	27.48
hera	(the) time (NOM)	5.78	6.51	17.09
normale	normal	5.47	7.90	18.58
mjafton	is enough	5.64	6.71	16.75
thotë	says	5.14	21.41	45.02
faktin	(the) fact (ACC)	5.29	10.91	24.11

shërbejë	serve (subjunctive)	5.79	5.61	14.72
zgjasë	last (subjunctive)	6.17	4.70	14.11
kuptohet	(is) understood	5.35	6.87	15.50
praktikë	practice	5.73	5.10	13.14
aspak	at all	5.19	7.77	16.59
rëndësi	importance	5.22	7.25	15.66
zbardhet	(is) whitened	6.63	3.90	13.71
varet	depends	5.45	5.84	13.67

Table 49: Top thirty collocations for *kjo*

In Table 49 are observed some of the same facts seen in the collocations of other *k(ë)-* words. There is a first person verb and pronoun, several verbs in subjunctive mood, and mostly passive verb constructions. While *ajo* has as one of its top collocations the *yesterday's one*, *kjo* has *the today's one*. A constant in this class of words remains *pikërisht*.

The demonstratives atë *and* këtë

The demonstratives *atë* and *këtë* have the attributes of singular and accusative. These forms are shared by both masculine and feminine. The distal *atë* is predominantly a pronoun. The proximal *këtë* is heavily unbalanced in favor of the usage as a determinative. The proximal also has a much larger number of occurrences.

	pronoun	determinative
atë	7,972	3,256
këtë	8,816	54,156

Table 50: Pronominal and determinative usage of *atë* and *këtë*

As discussed in previous chapters, the heavy unbalanced usage numbers for *këtë* are due to its anaphoric/cataphoric role in discourse. The examples extracted randomly from the 62972 concordances of *këtë* illustrate well the phoric role in both determinative and pronominal functions. In each of the five examples, *këtë* is referring to entities inside the discourse.

...rikthimin pikërisht të socialistëve në pushtet. E Meta vetë	këtë	e ka artikuluar shpesh. [33]
...ai më dha	këtë	mesazh: Edhe biznesmenët duhet të jenë…
...të shkruaja një artikull për të dhënë, pikërisht	këtë	mesazh, vendosa të riprodhoj tregimin e tij...
...mendonte për rritjen e kapacitetit të saj. Në	këtë	moment, një oportunitet...
...do të kontraktonin me të për prodhim gruri. Në	këtë	mënyrë ai do ta eliminonte...

Concordance 7: Random concordances illustrating *këtë* in its phoric roles

The fact that the most frequent form in the pair refers mainly to discourse entities is another pattern emerging from this kind of analysis.

The semantic analysis of the top collocations and tenses continues to sustain the patterns created until now.

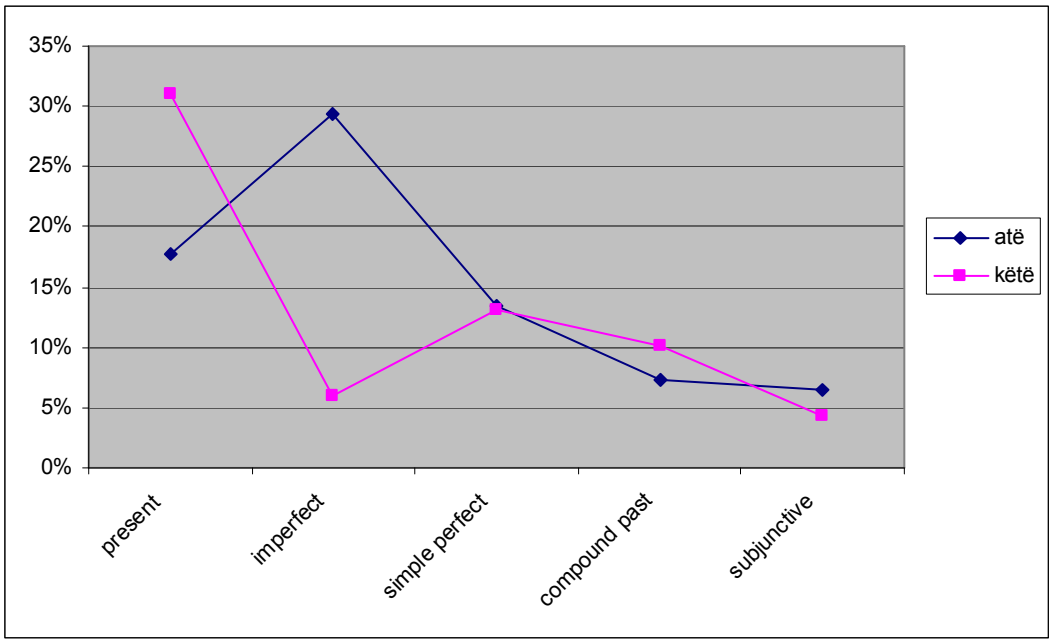


Figure 18: Distribution of tenses for *atë* and *këtë*

Present tense collocates with *këtë* the same way imperfect collocates with *atë* following the already established pattern by the other *a-* and *k(ë)-* words. The other two pasts

[33] ...precisely the return of the socialists in power. And Meta himself has articulated **this** often…
...he gave me **this** message: Even businessmen should…
...I would write an article that would have given precisely **this** message, I decided to reproduce his story…
...he was thinking about the growth of its capacity. At **this** moment, an opportunity…
...they would contract with him for wheat production. In **this** manner, he would eliminate…

and subjunctive have almost same values and therefore cannot affect the overall differences between present and past in general.

While the pattern established by tense still holds, the following two tables show no differences in the distribution of first or second person (either verbal or pronominal) which are equally frequent for both *atë* and *këtë*. This change coincides with huge imbalance in favor of *këtë*.

Collocate	English	MI-score	T-score	Z-score
njëjtë	same	7.26	9.62	42.03
ngjashëm	similar	7.52	7.77	37.25
quajti	called	6.75	8.17	29.94
ta	[clitic]	5.87	43.34	117.12
im	my	5.82	11.42	30.30
quan	calls	6.27	6.34	19.67
bir	son	6.34	5.98	19.04
pason	passes	7.29	4.03	17.82
ndodhte	happens (imperfect)	5.89	7.59	20.68
quhet	(is) called	5.74	8.51	22.00
ndodhi	happened	5.61	12.15	30.01
njihja	(I) know (imperfect)	6.32	4.32	13.63
meritonte	deserves (imperfect)	6.86	3.72	14.18
krahasohet	(is) compared	6.94	3.62	14.19
ngjashme	similar	5.80	5.49	14.50
dënoi	condemned	5.94	4.68	12.97
thashë	(I) said	5.42	7.63	17.65
shoqëronte	accompanies (imperfect)	6.17	3.77	11.30
cilësuar	qualified	5.36	7.07	16.03
ndodhur	happened	5.20	13.58	29.12
quante	call (imperfect)	6.31	3.60	11.33
shoqëron	accompany	6.10	3.75	10.97
ndodh	happens	5.25	7.89	17.19
meriton	deserves	5.76	4.34	11.29
kisha	(I) have (imperfect)	5.16	10.54	22.31
donim	(we) want (imperfect)	6.38	3.38	10.90
quajtën	called	6.26	3.47	10.75
përafërt	approximate	6.52	3.16	10.71
yt	your	5.86	3.76	10.12
shndërrojë	transform (subjunctive)	7.53	2.87	13.77

Table 51: Top thirty collocations for *atë*

In the list of the top thirty collocations for the demonstrative *atë*, a prominent part is occupied by adjectives and verbs used for expressing similarity and comparisons ('same,' 'similar [masculine],' 'compared,' 'similar [feminine],' 'qualified,' 'approximate,' and 'transform').

Collocate	English	MI-score	T-score	Z-score
enkas	on purpose	6.24	6.69	20.54
pikërisht	precisely	5.36	23.29	52.84
komentoni	(you) comment	6.92	5.37	20.90
implikuar	implicated	5.57	8.19	19.96
ta	[clitic]	5.13	49.37	103.24
lidhje	connection	5.07	21.29	43.70
pjesëmarrës	participant	5.38	6.80	15.49
zgjidhë	resolve (subjunctive)	5.55	5.50	13.32
përfshihen	(are) included	5.19	7.66	16.38
lidhur	connected	4.97	18.66	36.90
shkollor	scholarly	5.33	5.94	13.31
përfshira	included	5.82	4.54	12.04
zgjidhur	resolved	5.03	8.37	16.88
pranishëm	present	4.96	9.39	18.51
përdorën	used	5.64	4.60	11.50
akademik	academic	5.47	4.92	11.57
gisht	finger	5.32	5.25	11.74
ftuar	invited	5.00	7.90	15.81
ilustruar	illustrated	6.34	3.82	12.17
dëshiroj	(I) desire	5.33	5.07	11.36
them	(I) say	4.82	14.06	26.43
quajti	called	5.06	6.18	12.64
argumentuar	reasoned	5.26	4.68	10.23
bëni	(you) do	4.99	5.90	11.77
bukurie	beauty	5.75	3.81	9.88
thoni	(you) say	4.92	6.20	12.05
vërtetojnë	certify	5.36	4.34	9.82
vërtetoj	(I) certify	6.65	3.32	11.77
bëjmë	(we) do	4.72	9.86	17.87
shfrytëzoj	(I) exploit	7.17	3.13	13.30

Table 52: Top thirty collocations for *këtë*

The word *pikërisht* 'precisely' appears again at the top of the *k(ë)-* word collocations supported by *pranishëm* 'present'. The words *enkas* 'on purpose' and *dëshiroj* 'I desire' could both be considered as contributing to the subjective point of view of the demonstrative *këtë*.

By contrast, the various terms of comparison, as well as the several forms of 'happens' and 'deserves' in the collocations of *atë* support a more objective point of view. Chafe's point of view (1985) that nominal referents of demonstrative pronouns are often missing in spoken discourse due to the faster production and the lack of editing associated with speech cannot be supported by our corpus of written texts. Our corpus shows that Albanian demonstratives, in their pronominal roles cannot be errors. If that were true, one of the more frequent collocations would have been the word *gjë* 'thing' which instead has the following collocation scores with *këtë*: MI=2.61, T=-4.90, Z=-4.29. The negative T and Z-scores show that these words are dissociated.

As a conclusion, the effect of the high frequency of *këtë* shows in the smaller differentiation between the distal and the proximal. Frequency appears to remove markedness from *këtë* allowing for a heavy phoric role. Still some semantic features of subjectivity and therefore closeness to the *origo* remain.

The demonstratives ata *and* këta

The demonstratives *ata* and *këta* have the attributes of masculine and plural. Their case is either nominative or accusative. As illustrated in Table 53, *ata* is predominantly a pronoun which makes it mostly an anaphora. The word *këta* instead, does not have the imbalance of *ata* being split between its determinative and pronominal role on 2.5-to-1 ratio.

	pronoun	determinative
ata	21,891	912
këta	1,291	3,319

Table 53: Pronominal and determinative usage of *ata* and *këta*

This data coincides with the collocation data. Indeed, *ata* has a verb in 85% of the collocations while *këta* maintains the same ratios as the other demonstratives. All top thirty

words for *ata* are verbs in the third person plural agreeing with the pronoun. The same is observed for the rest of the verbs collocated with *ata*. Apparently, this high percentage of verbs collocated with *ata* causes the graph in Figure 19 below to show isomorphic lines in the distribution of tenses between *ata* and *këta*.

As for *këta*, in its list of top collocated words is found *pikërisht* 'precisely' again. While verbs agree in the third person plural and their majority is in present tense, there are also verbs and pronouns in the first person as well as other expressions referring to some kind of proximity with the *origo* such as *të sotme* 'of today' and *të fundit* 'the last.'

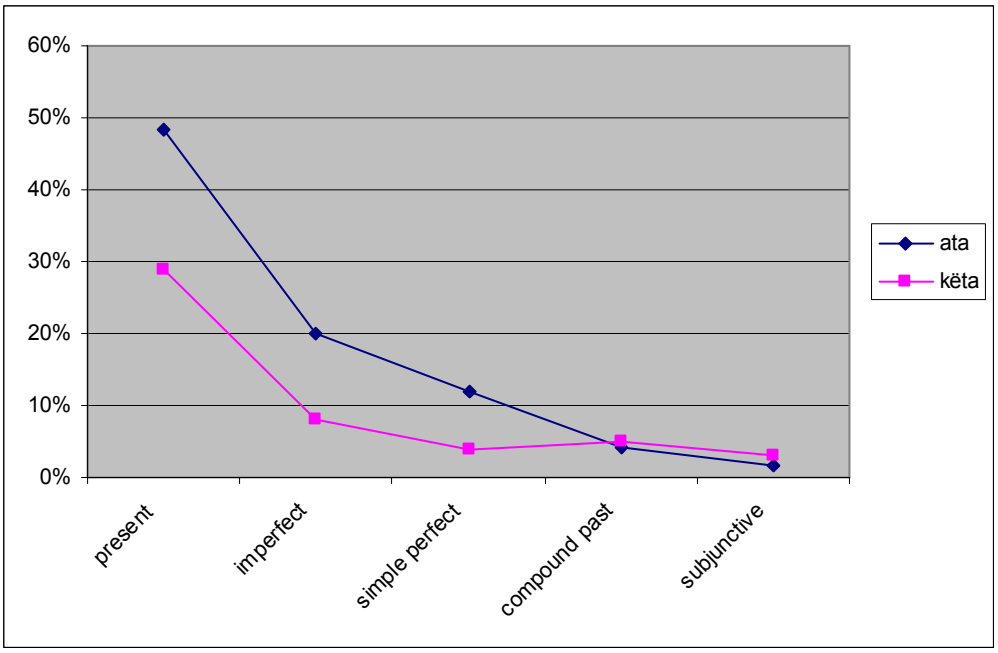


Figure 19: Distribution of tenses between *ata* and *këta*

Observing the lines in Figure 19 as well as the list of collocations for *këta*, we can conclude that *këta* is defined by its collocations in the same way as the other *k(ë)-* words. The word *ata* on the other hand requires further attention through the analysis of its concordances.

...Fshatarët dhe, përgjithësisht,	ata	që i takojnë një brezi...[34]
...nëse atje nuk jetojnë…	ata	mund të mbjellin...
...njerëzve të mbledhur në sallë. …	ata	shfaqnin… një dashuri ose...
...i përshkojnë njerëzit, kur	ata	bashkohen dhe formojnë turmën?
...në vendin e tyre. Por	ata	nuk mund të mos e kuptojnë…

Concordance 8: Random concordances for *këta*

In each of the concordances above and in most of the 22803 lines generated by *ata*, the pronoun is used as an anaphora whose antecedent is a collective noun. By referring to collective nouns even in cases when there is no clear antecedent, the pronoun behaves like the generic *they* in English. This usage also explains the counterintuitive dominance of present tense in the collocations of the distal form.

The demonstratives ato *and* këto

The demonstratives *ato* and *këto* have the attributes of feminine, plural and either nominative or accusative case. The distribution of the demonstratives among the pronominal and determinative roles is more balanced for *ato* than for *këto* reinforcing the idea that *k(ë)-* words are mostly determinatives while *a-* words are mostly pronouns.

	pronoun	determinative
ato	11,581	4,284
këto	2,581	18,927

Table 54: Pronominal and determinative usage of *ato* and *këto*

The distributions in Table 54 and, therefore, their functionality, influence the distribution of verbal tenses as well. As happened with the masculine forms, the lines describing the distribution of tenses in the graph below are almost isomorphic. The biggest

[34] …The villagers and, in general, those that belong to a generation…
…if over there don't live… they can cultivate…
…of the people assembled in that room. …they expressed… love or…
…goes through people, when they unite and form a mob?
…in their country. But they can't not understand…

differences are on imperfect where *ato* gets twice as much occurrences compared to *këto* and corresponds with previous distributions.

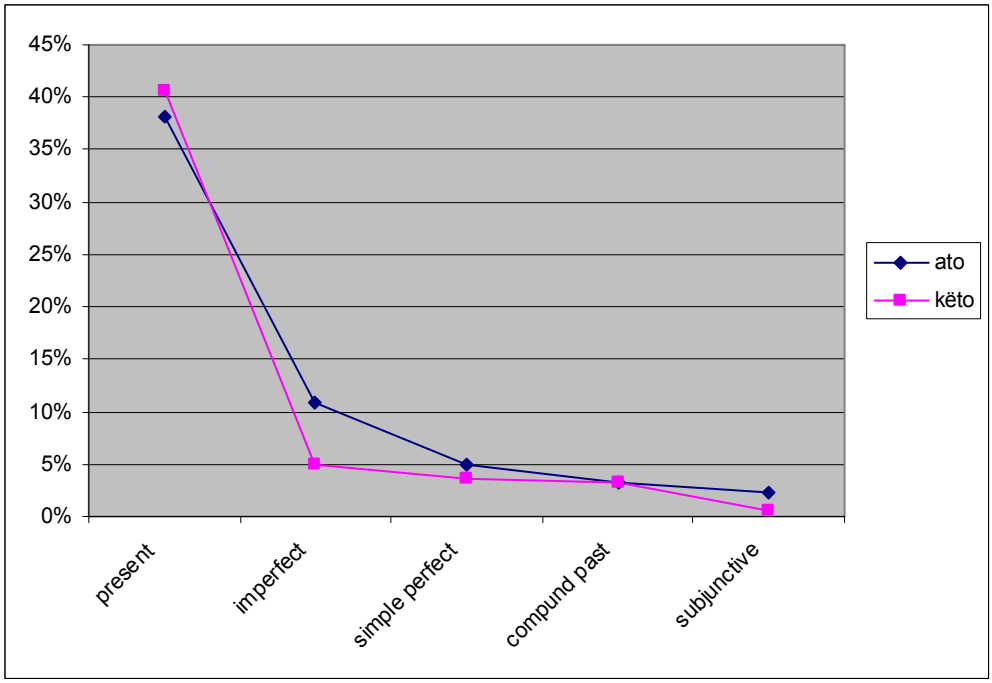


Figure 20: Distribution of tenses between *ato* and *këto*

The words *pikërisht* 'precisely' and *sotmet* 'of today' remain in the list of top collocations for the *k(ë)-* word. While *ato* has in its list of top collocations expressions of similarity such as *ngjashme* 'similar,' *identike* 'identical,' *njëjta* 'same,' *këto* has a concentration of calendar/time expressions such as *vite* 'years,' *vjetët* '(the) years,' *netë* 'nights,' *treja* 'all three,' *fundit* 'last,' etc. As with the adverbs of place, the *kë-* word refers to a time close to the *origo*. This might be caused by the corpus which contains a large amount of journalistic sources. Since journals write mostly about current events, it makes sense to have these collocations. As for the similarity expressions collocated with the distal, they can be explained with need of the speaker/writer to identify something unknown (referred to by *ato*) with features from a known object.

The demonstratives atij *and* këtij

The feature attributes for *atij* and *këtij* are masculine, singular, and either dative, genitive or ablative.

	pronoun	determinative
atij	2,869	1,350
këtij	912	21,897

Table 55: Pronominal and determinative usage of *atij* and *këtij*

In the table above we notice a heavy imbalance in favor of the determinative function of *këtij*. While the graphical representation for the tense distribution of *atij* remains more or less similar to that of other *a-* words, *këtij* appears as if it belonged to a completely different class.

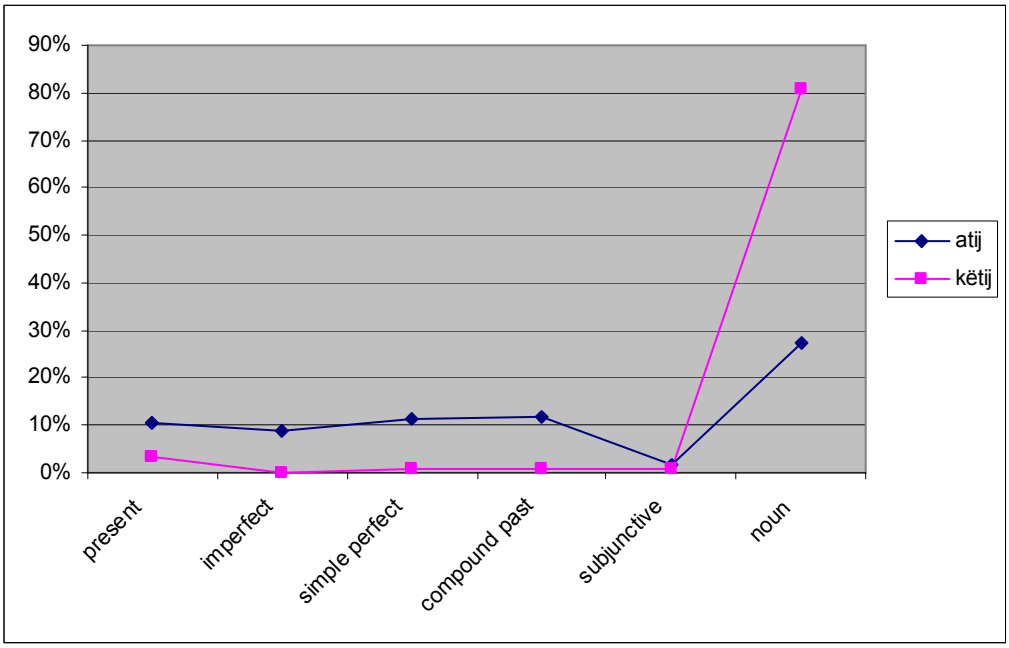


Figure 21: Distribution of tenses between *atij* and *këtij*

To explain this change in *këtij*, we added a data point with the percentage of nouns in the demonstratives' list of collocations. With more than 80% of the collocations being nouns, we can also explain the large imbalance in favor of the determinative role for *këtij*.

By drilling further into the list of collocations for *këtij*, the glaring categories are those of calendar expressions and nouns derived from verbs. The top fifty collocated words include all the twelve names of months, seasons, and several other time periods like semester, date, and monthly. The word *pikërisht* is not anymore a top collocation for the *k(ë)-* word in this case.

While verbs are almost missing, verb derived nouns make 26% of the top collocations and include nouns like *realizimin* 'realization,' *zgjidhjen* 'solution,' *fillimin* 'beginning,' *miratimi* 'approval,' *zbatimi* 'implementation,' and *krijimi* 'creation' as in the concordances below.

...ndërtimit të	këtij	aksi rrugor…[35]
...zgjidhjen e	këtij	problemi…
...dhënien e	këtij	dokumentacioni…
...fshehjes ose zhdukjes së	këtij	dokumentacioni…
...modernizimin e	këtij	sektori…

Concordance 9: Verb derived nouns near *këtij*

The direct object of the nominalized verb becomes a genitive in Albanian (rendered with a prepositional phrase in the English translation in the footnote).

The genitive nouns determined by a demonstrative expose the nature of the determinatives in general. While the generic rule is that an adjective agrees in gender, number and case, the demonstrative determinative cannot be a genitive which is why it is not labeled as an adjective. Also, the demonstrative determinative, while it can appear either in the *a-* or

[35] …the construction of this route…
…the solution of this problem…
…the rendering of this documentation…
…the hiding or disappearance of this documentation…
…the modernization of this sector…

k(ë)- form, cannot be rendered as the unprefixed form *tij* (or *saj* / *të* / *ta* / *to* / *tyre*) which assume only the role of a pronoun.

The demonstratives asaj *and* kësaj

The feature attributes for *asaj* and *kësaj* are feminine, singular, and either dative, genitive or ablative.

	pronoun	determinative
asaj	2,762	1,842
kësaj	1,604	18,444

Table 56: Pronominal and determinative usage of *asaj* and *kësaj*

In the table above, the distribution of the pronominal and determinative roles follows the same pattern as in *atij/këtij* distributions table. The *a-* word has a balanced distribution while the *k(ë)-* word maintains its predominantly determinative functionality. This distribution is also reflected in the grammatical categories of their corresponding collocations.

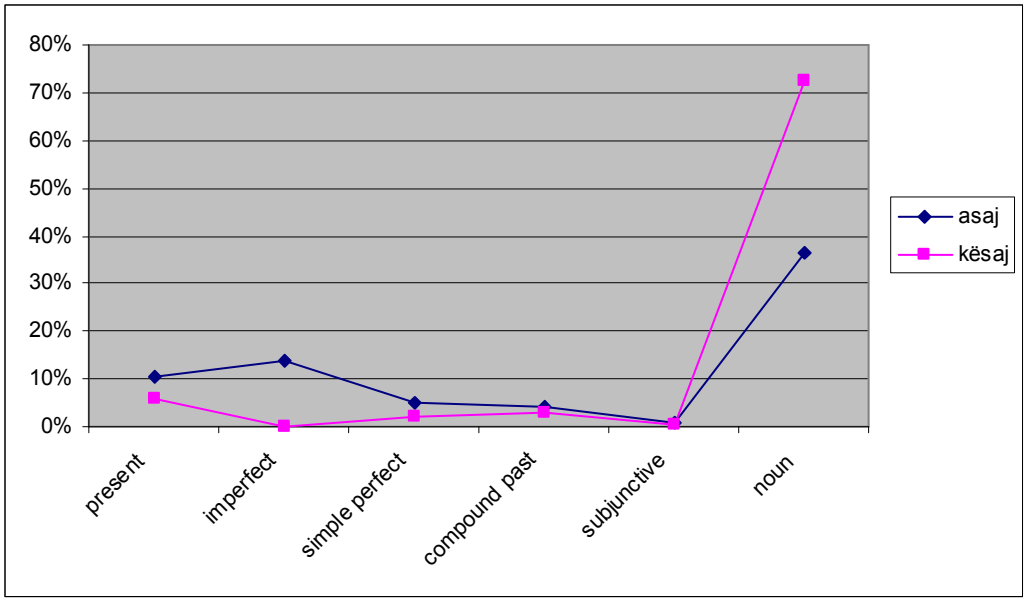


Figure 22: Distribution of tenses between *asaj* and *kësaj*

As with the masculine forms, the collocations of *këtaj* are mostly nouns. The graph still shows the incompatibility between a *k(ë)-* word and imperfect which in this case is absolute. Observing the list of collocations for *këtaj*, the incompatibility with the imperfect tense is reinforced by the top collocated word *sot* 'today.'

The demonstratives atyre *and* këtyre

The grammatical features attributes for *atyre* and *këtyre* are plural, and either dative, genitive or ablative. This form does not differentiate between masculine and feminine.

	pronoun	determinative
atyre	4,149	1,170
këtyre	194	9,492

Table 57: Pronominal and determinative usage of *atyre* and *këtyre*

The table above illustrates the same imbalances as the other pairs. Upon further analysis of the proximal demonstratives, they determine a genitive noun 62% of the cases, an ablative in 34% of the cases and dative the rest (4%). The distal determinative has a more balanced distribution with 53%, 27% and 20% respectively. When in pronominal roles, the distal forms are in 62% of the cases dative, 31% genitive and 7% ablative. Proximal forms in pronominal roles are very rare and are found mostly in dative and ablative. The more balanced distribution of cases for *atyre* suggests a larger variety of syntactic roles (indirect object being one of them.) This fact is also reflected in the distribution of tenses.

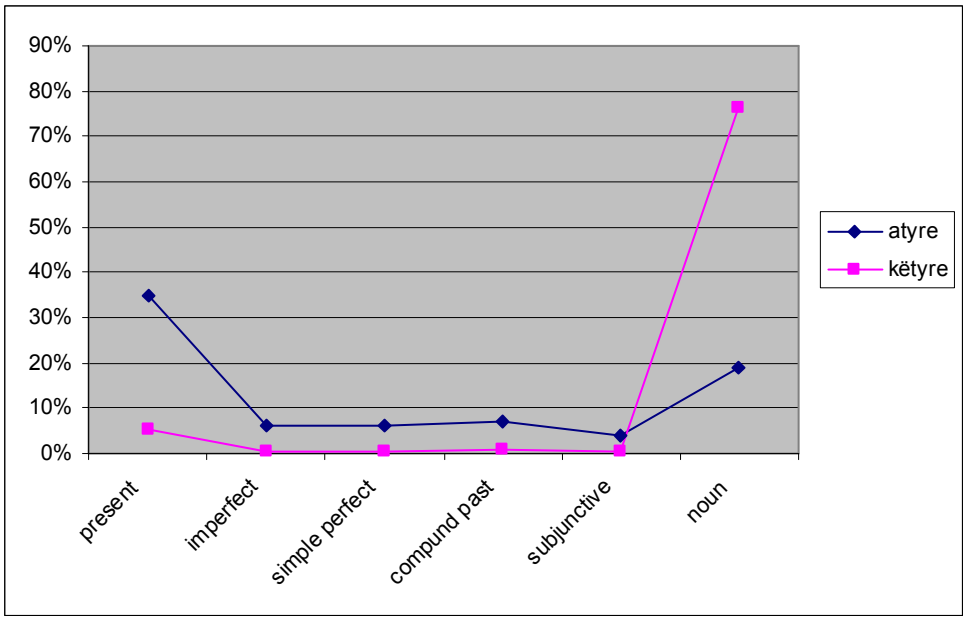


Figure 23: Distribution of tenses between *atyre* and *këtyre*

The graph for *këtyre* shows a very low number of verbs. The predominant function of this word is that of a determinative. However, even among the very limited number of verbs, present tense is dominant.

...letërsinë e krijuar gjatë	këtyre	12 vjetëve të fundit…[36]
...madje brënda	këtyre	ditëve të fundvitit…
...me politikanët shqiptarë gjatë	këtyre	ditëve…
...ishte strategjia e	këtyre	dhe këta kërkuan ta ruanin këtë...
…shumica e	këtyre	qytetarëve kanë humbur shumë…

Concordance 10: Random concordances for *këtyre*

As it can be seen in this random sample of five concordances, four of them are determinative, three of which determine the words *day* and *year*. This indeed is the pattern in the list of collocations: in the top twenty collocated words we find *vjetëve* 'years

[36] …literature created during these last 12 years…
…indeed between these end-of-the-year days…
…with the Albanian politicians during these days…
…it was the strategy of these (ones) and they tried to save this…
…most of these citizens have lost a lot…

GEN/DAT/ABL,' *vitet* 'years NOM/ACC,' *muajve* 'months GEN/DAT/ABL,' *javëve* 'weeks GEN/DAT/ABL,' and *ditëve* 'days GEN/DAT/ABL.'

The distal demonstrative instead has a higher number of verbal collocations but still atypical because of the higher ratio of present tense verbs.

...është edhe vendi i	atyre	njerëzve që nuk pajtohen...[37]
…një dekadë pas	atyre	zgjedhjeve, të zhgënjyer nga…
...u shpjegon	atyre	se shumë probleme vijnë…
...zgjedhjeve parlamentare në Maqedoni dhe	atyre	lokale në Kosovë…
...ata të votojnë kundër	atyre	, të cilët e sollën shtetin…

Concordance 11: Random concordances for *atyre*

The high number of present tense is explained by the very large collocation scores achieved by the word *që* 'that' (MI=3.57, T=10.39 and Z=12.78) as in the political slogan *t'u japim shpresë atyre që shpresojnë tek ne* 'let's give hope to those that have faith in us.' "Those that…" refers to the generic *they* discussed earlier in the chapter "The demonstratives *ata* and *këta*."

By collapsing the data from the three pairs studied last, and applying a collocation strength measurement that does not discriminate for frequency (combining Z and T scores with a log likelihood measurement) we come up with these lists of collocations that show the differences between the proximals and the distals under a different light.

[37] …is the country of those people that do not agree…
…one decade after those elections, disappointed by…
…explain to those (ones) that many problems come…
…parliamentary elections in Macedonia and those local in Kosovë…
…they vote against those who brought the government…

	atij	asaj	atyre	këtij	kësaj	këtyre
collocations with frequency bias	takon	quhet	ua	qëllimi	zbardhjen	vjetëve
	ia	ndodhi	duan	mars	veç	gjatë
	dha	takon	iu	fillim	përveç	viteve
	t	ndodhur	përveç	gjatë	gjatë	banorët
	iu	shqiptare	u	brenda	pas	përveç
	kërkuar	asaj	sidomos	sipas	menjëherë	ditëve
	dhënë	mes	numri	fundit	sot	fundit
	i	midis	t	fund	brenda	prej
	mes	ia	gjithë	pas	pranë	midis

Table 58: Collocations for the dative/genitive/ablative forms

In the table above, the first three columns, which contain the collocations of the distals, are dominated by dative clitics that are mandatory in Albanian and verbs that require an indirect object suggesting the mostly pronominal usage of the *a-* words. The other half of the table is dominated by temporal reference terms, ablative prepositions and nouns. These categories point to the mainly determinative role of the *k(ë)-* words. The expressions referring to time include *sot* 'today,' *menjëherë* 'immediately,' and *fundit* 'last,' all of which are expressions pointing near the NOW element of the *origo*.

Summary

The analysis of the Albanian demonstratives provides us with a new classification in which we count them by their lexical grammatical category. The total number of occurrences in the two categories taken together is very similar. As seen in Table 59, the differences are found in the distribution of the proximals and distals into one category or the other. The number of occurrences describes a situation where proximal demonstratives assume mainly the role of a determinative while the distal demonstratives assume the role of a pronoun.

demonstratives			
determinative	occurrences	pronoun	occurrences
këtë	54,156	ai	42,444
këtij	21,897	ata	21,891
këto	18,927	ajo	18,299
kësaj	18,444	kjo	12,672
kjo	13,727	ato	11,581
ky	12,426	këtë	8,816
këtyre	9,492	atë	7,972
ato	4,284	ky	5,076
këta	3,319	atyre	4,149
atë	3,256	atij	2,869
asaj	1,842	asaj	2,762
ajo	1,591	këto	2,581
atij	1,350	kësaj	1,604
atyre	1,170	këta	1,291
ata	912	këtij	912
ai	866	këtyre	194
total	167,659	total	145,113

Table 59: Distribution of determinatives and pronouns

The other feature of the demonstratives observed in the table above is that pronouns are mostly in the nominative case and therefore mostly in the role of a subject. The determinatives are mostly in accusative, dative and ablative: a fact that leads us to deduce that they determine mostly nouns in object and complement roles. Another fact discussed briefly in the chapter "The demonstratives *atij* and *këtij*" is the lack of the genitive case when the demonstrative is a determinative. The lack of genitive leads to the conclusion that the determinative, even though it determines nouns in all cases[38], has only three forms as in the examples for *libër* 'book' and *fletore* 'notebook' in Table 60. It is also observed that the plural

[38] The construct determinative + genitive/ablative requires first the attachment of the determinative to the nominal component and then the attachment of the genitive particle or ablative preposition. It appears that determinatives, even though they are bound to the deictic particles, they still maintain certain features of the Albanian definite articles derived from the old demonstratives.

determinative cannot determine the indefinite form ending in –*sh* (*librash*, *fletoresh*) which would be possible only with the very rare forms *asi* and *kësi* as in *asi/kësi librash/fletoresh* 'of that/this kind of books/notebooks.' When the demonstrative is a pronoun it does have all cases.

	singular		plural	
nominative	ai/ky libër	ajo/kjo fletore	ata/këta libra	ato/këto fletore
definite	ai/ky libri	ajo/kjo fletorja	ata/këta librat	ato/këto fletoret
accusative	atë/këtë libër/fletore		ata/këta libra	ato/këto fletore
definite	atë/këtë librin/fletoren		ata/këta librat	ato/këto fletoret
dative	atij/këtij libri	asaj/kësaj fletore	atyre/këtyre librave/fletoreve	
definite	atij/këtij librit	asaj/kësaj fletores		
genitive (i, e, të, së)	atij/këtij libri	asaj/kësaj fletore	atyre/këtyre librave/fletoreve	
definite	atij/këtij librit	asaj/kësaj fletores		
ablative (prej, përveç, etc.)	atij/këtij libri	asaj/kësaj fletore	atyre/këtyre librave/fletoreve	
definite	atij/këtij librit	asaj/kësaj fletores		

Table 60: Inflection of determinatives

Even though both proximal and distal demonstratives can take a determinative role, the proximals dominate this category numerically. From the examples discussed in the previous chapters, we observed that the most frequent forms, the proximals, have lost their distance meaning and have assumed mostly a definiteness trait. When distals are used as determiners, the semantic component that marks indefiniteness and loose connectivity to the *origo* is very strong. The inverse is true for pronouns: the use of distals as anaphora weakens the distance element. When a proximal is used as a pronoun (a much rarer event), its element of definiteness and closeness to the *origo* (either endophoric or exophoric) is very strong. In other languages (Biber, 1988), anaphoric pronouns, often called 3rd person personal pronouns, mark relatively inexact reference to persons outside of the immediate interaction. Third person pronouns co-occur frequently with past tense forms, as markers of narrative, reported styles. Albanian, with its use of the proximal pronoun in anaphoric roles makes this

generalization more difficult, thus leading us to the discussion of the personal pronouns in Albanian.

As in many other languages, 1st and 2nd person pronouns are proper deictics. The third person has a dual anaphoric and deictic nature making it hard to be classified under one or the other. If the pronoun is anaphoric, according to most grammars of Albanian, it is classified as a 3rd person personal pronoun. If it is deictic, it is categorized as a demonstrative pronoun. While diachronic analysis provides a good explanation of how the demonstratives evolved in Albanian, synchronic analysis offers no clear division between personal and demonstrative pronouns. This new quantitative dimension moves us towards a better definition of personal and demonstrative pronouns. On the one hand, these pronouns do keep a high level of association with their corresponding deictic family. On the other hand, both groups find themselves associated with words such as verbs that agree with the analyzed pronoun and that would fit in the same syntactic role. The main conclusions reached by this analysis are:

1. Albanian demonstrative pronouns maintain their deictic functionality for both endophoric and exophoric references.

2. Pronouns that contain *a-*, *k(ë)-* (or in some cases neither) are syntactically interchangeable.

3. Collocational analysis provides additional arguments for determining the syntactic unity of demonstratives while maintaining their deictic differences.

Combining insights from diachronic studies with synchronic and quantitative studies, the implications that emerge include the primacy of deixis in the development of the

pronominal systems in general. Albanian's lack of third person proper shows a path of language evolution that maintains its deictic elements both in referential and anaphoric functions. While both *a-* and *kë-* pronouns play the role of what is called third person they preserve their deicticity. The *ø-* pronouns, never appearing without a preposition, etymologically belong to the same demonstrative paradigm. Functionally, prepositions neutralize the need for deictic prefixes allowing them to disappear in some cases. The continuum between anaphoric and deictic functions does not include a cusp that divides the two. The lack of a 3rd person personal pronoun form classifies the Albanian language as a two-person language in Bhat's (2004) taxonomy.

Epilogue

> 'When *I* use a word,' Humpty Dumpty said, 'it means just what I choose it to mean – neither more nor less.'
>
> 'The question is,' said Alice, 'whether you *can* make words mean so many different things.'
>
> 'The question is,' said Humpty Dumpty, 'which is to be master – that's all.'
>
> Alice was too much puzzled to say anything; so after a minute Humpty Dumpty began again. 'They've a temper, some of them – particularly verbs: they're the proudest – adjectives you can do anything with, but not verbs – however, *I* can manage the whole lot of them! **Impenetrability**! That's what *I* say!'
>
> Lewis Carroll
> *Alice's Adventures in Wonderland*
> *Through the Looking Glass*
> Oxford, 1865 and 1872
> 2007 Blackwell Alice's Day Edition

It is verbs that have the most features and this makes them the most interesting from the gravitation point of view. The *impenetrability* of these features goes beyond the arbitrarity of the linguistic sign: it is not the words but their features that define what goes around. Like planets, we found out that these features have a clearly defined gravitational pull. The words that we analyzed find themselves entangled in bundles of features sometime coming from different components. It is the relationships created by these bundles that end up defining the words themselves.

An Albanian deictic is realized historically in two forms. Each of these forms – one headed by the prefix *a* and the other by the prefix *k(ë)* – is found in the orbits of words that, once analyzed, expose the meaning of the deictics. Linguistic literature calls them distals and proximals but distance is only a small component in the meanings of these words. As we found out, distance becomes significant only in the rarest of occurrences.

The main semantic components of Albanian deictics are defined through their relationship with the grammatical categories of tense, person, and definiteness, with the focus on the speaker during his act of speech or with the *origo*, with the familiarity of the subject matter, and other features as well. These quantifiable relationships establish two syntactically similar categories which have significant differences semantically.

The Albanian deictic system is binary. The words beginning with *a-* circulate in the orbits of words that refer to undefined situations, in not well defined times and in not well defined locations. Using terms borrowed from geometry, we would be in presence of unbounded sets, lines, or planes. The words starting in *k(ë)-* instead, circulate in the orbits of words that refer to well defined situations, happening in specific times and places. Returning to the borrowed terminology, we are here in presence of closed bounded sets.

Bibliography

AA. VV. [FGSS] 1980. *Fjalor i Gjuhës së Sotme Shqipe*. Tiranë: Akademia e Shkencave.

AA. VV. [FSS] 1984. *Fjalor i Shqipes së Sotme*. Tiranë: Akademia e Shkencave.

Abir, E., S. Klein, D. Miller, and M. Steinbaum. 2002. Fluent Machines' EliMT System. In *Fifth Conference of the Association for Machine Translation in the Americas*, *AMTA-2002*, Tiburon, CA. 216-219.

Agalliu, F., E. Angoni, S. Demiraj, A. Dhrimo, E. Hysa, E. Lafe, E. Likaj. 2002. *Gramatika e Gjuhës Shqipe, Vol. 1*. Tiranë: Akademia e Shkencave.

Altimari, F. and F. De Rosa. 1995. *Testi folclorici di Falconara Albanese*. Rende: Diadia.

Barnbrook, G. 1996. *Language and Computers*. Edinburgh: Edinburgh University Press.

Benedetti, M. and D. Ricca. 2002. The systems of deictic place adverbs in the Mediterranean: some general remarks. In Ramat, P. and T. Stolz (eds.). *Mediterranean languages: Papers from the MEDTYP workshop, Tirrenia, June 2000*. Bochum: Universitätsverlag Dr. N. Brockmeyer. 13-32.

Benor, S. and R. Levy. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language* 82. 233-278.

Benveniste, E. 1962. *Hittite et Indo-Européen*. Paris: Librairie Adrien Maisonneuve.

Benveniste, E. 1966. *Problèmes de linguistique générale*. Paris: Gallimard.

Bertinetto, PM. 1986. *Tempo, Aspetto e Azione nel verbo italiano. Il sistema dell'indicativo*. Firenze: Accademia della Crusca.

Bertinetto, PM. 1995. Compositionality and non-compositionality in morphology. In Dressler W. and C. Burani. *Cross-Disciplinary Approaches to Morphology*. Wien: Österreichische Akademie der Wissenschaften. 9-36.

Bertinetto, PM. 2001. On a Frequent Misunderstanding in the Temporal-Aspectual Domain: The 'Perfective - Telic' Confusion. In Cecchetto, C., G. Chierchia and M. Guasti (eds.). *Semantic Intefaces: Reference, Anaphora and Aspect*. Stanford: CSLI Publications. 177-210.

Bhat, D. N. S. 2004. *Pronouns*. Oxford: Oxford University Press.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., S. Conrad and R. Reppen. 1998. *Corpus Linguistics*. Cambridge: Cambridge University Press.

Bokshi, B. 2004. *Për Vetorët e Shqipes*. Prishtinë: Akademia e Shkencave dhe e Arteve e Kosovës.

Brinton, L. and E. Traugott. 2005. *Lexicalization and Language Change*. Cambridge: Cambridge University Press.

Busa, R. 1987. *Fondamenti di informatica linguistica*. Milano: Vita e Pensiero.

Bybee, J. and P. Hopper. 2001. *Frequency and the emergence of linguistic structure*. Amsterdam/Philadelphia: John Benjamins.

Bybee, J. 2006. From Usage to Grammar: the Mind's Reponse to Repetition. *Language* 82. 711-733.

Bühler, K. 1934/1990. *Theory of Language: the representational function of language*. Translated by Goodwin, D. Amsterdam & Philadelphia: John Benjamins.

Çabej, E. 1976a. *Studime Gjuhësore I, Studime Etimologjike në Fushë të Shqipes, A-O*. Prishtinë: Rilindja.

Çabej, E. 1976b. *Studime Gjuhësore II, Studime Etimologjike në Fushë të Shqipes, P-ZH*. Prishtinë: Rilindja.

Çabej, E. 1976c. *Studime Gjuhësore III*. Prishtinë: Rilindja.

Çabej, E. 1977. *Studime Gjuhësore IV*. Prishtinë: Rilindja.

Çabej, E. 1986. *Studime Gjuhësore VII*. Prishtinë: Rilindja.

Çabej, E. 1994. *Për Gjenezën e Literaturës Shqiptare*. Tiranë: MÇM.

Camaj, M. 1977. *Die albanische Mundart von Falconara Albanese in der Provinz Cosenza*. München: Dr. Dr. Rudolf Trofenik.

Carbonell, J., S. Klein, D. Miller, M. Steinbaum, T. Grassiany and J. Frey. 2006. Context-Based Machine Translation. In *AMTA 2006*. Boston. 19-28.

Chafe, W. 1985. Linguistic differences produced by differences between speaking and writing. In Olson, D.R., N.Torrance, and A. Hildyard (Eds.), *Literature, language, and learning: The nature and consequences of reading and writing*. New York: Cambridge University Press. 105-123.

Champseix, J. P. 2006. Entre les lignes, la tyrannie. In *Cahiers balkaniques n° 35 : Linguistique, littérature, civilisation*. Paris : INALCO. 73-85.

Choueka, Y. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *RIAO'88 Conference Proceedings*. Cambridge: MIT. 609-623.

Church, K. and P. Hanks. 1991. Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics 16(1)*. 22-29.

Church, K., W. Gale, P. Hanks, and D. Hindle. 1991. Using Statistics in Lexical Analysis. In Zernik, U. (ed.). *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Hillsdale: Lawrence Erlbaum. 115-164.

Church, K. and R. Mercer. 1993. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. In *Computational Linguistics 19(1)*. 1-24.

Coseriu, E. 1997. *Linguistica del Testo*. Roma: Carocci.

De Rosa, F. 2003. *I Canti di Serafina Thopia: Edizione critica delle tre versioni dell'opera (1839, 1843, 1898)*. Rende: Luigi Pellegrini Editore.

Demiraj, S. 2002. *Gramatikë Historike e Gjuhës Shqipe*. Tiranë: Akademia e Shkencave.

Dhrimo, A., E. Angoni, E. Hysa, E. Lafe, E. Likaj, F. Agalliu, et al. 1986. *Fonetika dhe Gramatika e Gjuhës së Sotme Shqipe: Morfologjia*. Tiranë: Akademia e Shkencave.

Di Giovine, P. 1990. *Studio sul perfetto indoeuropeo*. Roma: Dipartimento di Studi Glottoantropologici dell'Università di Roma "La Sapienza."

Diessel, H. 1999. *Demonstratives: Forms, Function, and Grammaticalization*. Amsterdam and Philadelphia: John Benjamins.

Durham, E. 2000. *High Albania: a Victorian Traveller's Balkan Odyssey*. London: Phoenix Press. (Original work published 1909).

Fillmore, C. 1982. Towards a Descriptive Framework for Spatial Deixis. In Jarvella, R. J. and W. Klein (ed.) 31-59.

Firth, J. 1957a. A synopsis of linguistic theory 1930–1955. In Special Volume of the Philological Society, *Studies in Linguistic Analysis*. Oxford: Basil Blackwell. 1–32.

Firth, J. 1957b. *Papers in Linguistics: 1934-1951*. London: Oxford University Press.

Firth, J. 1968. Descriptive linguistics and the study of English. In Palmer, F. (ed.) *Selected Papers of J. R. Firth – 1952-1959*. Bloomington and London: Indiana University Press. 96-113.

Friedman, V. 1981. Admirativity and Confirmativity. *Zeitschrift fur Balkanologie* 17: 1. 12–28.

Friedman, V. 1994. Ga in Lak and the Three "There"s: Deixis and Markedness in Daghestan. In Aronson, H. (ed.). *NSL.7: Linguistic Studies in the Non-Slavic Languages of the Commonwealth of Independent States and the Baltic Republics*. Chicago: Chicago Linguistic Society. 79-93.

Friedman, V. 1997. One Grammar, Three Lexicons: Ideological Overtones and Underpinnings in the Balkan Sprachbund. In *CLS 33 Papers from the 33rd Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society. 23-44.

Friedman, V. 2001. Hunting the Elusive Evidential: The Third-Person Auxiliary as a Boojum in Bulgarian. In Friedman, V. and D. Dyer (eds.) *Of All the Slavs My Favorites: In Honor of Howard I. Aronson*. *Indiana Slavic Studies* 12: 1-28.

Friedman, V. 2003. Vendi i gegnishtes në gjuhën shqipe dhe në Ballkan. *Phoenix, Nr. 1-6 (33-38)*. 40-56.

Friedman, V. 2004. *Studies on Albanian and Other Balkan Languages*. Pejë: Dukagjini.

Gahl, S. and S. Garnsey. 2004. Knowledge of Grammar, Knowledge of Usage. *Language* 80. 748-775.

Gawrych, G. 2006. *The Crescent and The Eagle - Ottoman Rule, Islam and the Albanians, 1874-1913*. London: I.B.Tauris. 181-182

Greenberg, J. 1966. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Greenberg, J. (ed.). *Universals of Language* (2nd ed.). Cambridge: MIT Press. 73-113.

Gries, S. 2006. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In: Gries, S. T. and A. Stefanowitsch (eds.). *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*. Berlin, Heidelberg, New York: Mouton de Gruyter (TiLSM 172). 57-99.

Hagège, C. 1992. Le système de l'anthropophore et ses aspects morphogénétiques. In Morel, M-A and L. Danon-Boileau (eds.). *La deixis: Colloque en Sorbonne (8-9 juin 1990)*. Paris: Presses Universitaires de France.115-123.

Halliday, M. A. K. and R. Hasan (1976). *Cohesion in English*. London: Longman.

Hamp E. 1993. *Il sistema fonologico della parlata di Vaccarizzo Albanese*. Rende: Università della Calabria. Italian translation of Hamp, E. 1954.*Vaccarizzo Albanese Phonology: The Sound-System of a Calabro-Albanian Dialect*.

Haspelmath, M. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42. 25-70.

Jarvella, R. and W. Klein (eds.) 1982 *Speech, Place, and Action: Studies of Deixis and Related Topics*. New York: John Wiley & Sons.

Jurafsky, D., A. Bell, M. Gregory, and W. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee and Hopper (2001:229-254).

Kallulli, D. 2001. Direct Object Clitic Doubling in Albanian and Greek. In Rivero, M. L. and A. Ralli (eds.). *Comparative Syntax of Balkan Languages*. Oxford: Oxford University Press. 127-160.

Kemp, A. 1987. The Tekhne grammatike of Dionysius Thrax. In Taylor, D. (ed.). *The History of Linguistics in the Classical Period*, Amsterdam. 169-189.

Конески, Б. 1981. *Граматика на Македонскиот Литературен Јазик*. Скопје: Наша Книга.

Krug, M. 2001. Frequency, iconicity, categorization: Evidence from emerging modals. In Bybee and Hopper (2001:309-335).

Kučera, H. 1982. Markedness and Frequency: A Computational Analysis. In Horecky, J. (ed.), *COLING 82*, Amsterdam: North-Holland Publishing Company. 167-173.

Labov. W. 2002. Driving Forces in Linguistic Change. In *International Conference on Korean Linguistics, August 2, 2002, Seoul National University*. <http://www.ling.upenn.edu/~wlabov/Papers/DFLC.htm> accessed March 8, 2008.

Labov, W. 2007. Transmission and Diffusion. In *Language, Journal of the Linguistic Society of America*, *83(2)*. 344-387.

Lehmann, W. 1982. Deixis in Proto-Indo-European. In Tischler, J. (ed.). *Serta Indogermanica: Festschrift für Günter Neumann zum 60. Geburtstag*. Innsbruck: Institut für Sprachwissenschaft. 137-142.

Leteux, C. 2006. L'évolution du lexique albanais depuis l'instauration de la démocratie en Albanie. In *Cahiers balkaniques n° 35 : Linguistique, littérature, civilisation*. Paris : INALCO. 57-72.

Likaj, E. 1984. *Analogjia gramatikore në strukturën morfologjike të gjuhës shqipe*. Tiranë: Universiteti i Tiranës.

Likaj, E. 1997. *Format analitike në gjuhën shqipe*. Tiranë: SHBLU.

Manning, C. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Mason, O., 2000. *Programming for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mayerthaler, W. 1988. *Morphological naturalness*. Ann Arbor: Karoma.

Mioni, A. 2000. Le macrocause dei mutamenti linguistici e i loro effetti. In Cipriano, P., R. d'Avino and P. Di Giovine (eds.). *Linguistica Storica e Sociolinguistica – Atti del Convegno della Società Italiana di Glottologia (Roma, 22-24 ottobre 1998)*. Roma: Il Calamo. 123-162.

Muller, C. 1973. *Initiation aux méthodes de la statistique linguistique*. Paris: Hachette.

Murzaku, A. 1988. Si përdoret përemri vetor "unë" në gjuhën shqipe. In *Gjuha jonë*. 4. Tiranë: Akademia e Shkencave. 91-93.

Murzaku, A. 1989. Përemrat "ai" dhe "ky" në gjuhën letrare shqipe, *Studime Filologjike*, 1. Tiranë: Akademia e Shkencave. 33-43.

Murzaku, A. 1990. Referenzialità dei pronomi deittici impuri dell'albanese. (16th International Congress of Albanology, Palermo.) In *Quaderni del Laboratorio di Linguistica*, Scuola Normale Superiore, Pisa, 1991.

Murzaku, A. 2007. Does Albanian have a Third Person Personal Pronoun? Let's have a Look at the Corpus… In Fitzpatrick, E. (ed.). *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*. Amsterdam: Rodopi. 243-255.

Nakhleh, L., D. Ringe, and T. Warnow. 2005. Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages. In *Language, Journal of the Linguistic Society of America, 81(2)*. 382-420.

Newmark, L. 1998. *Albanian-English Dictionary*. Oxford: Oxford University Press.

Orel, V. 1998. *Albanian Etymological Dictionary*. Boston-Leiden: Brill.

Pederson, E. and D. Wilkins. 1996. *A Cross-Linguistic Questionnaire on Demonstratives, Manual for the 1996 Field Season*. Nijmegen: Cognitive Anthropology Research Group, Max Planck Institute for Psycholinguistics.

Plank, F. and W. Schellinger. 1997. The uneven distribution of genders over numbers: Greenberg Nos. 37 and 45. *Linguistic Typology* 1. 53-101.

Plank, F. 1998. *Das Grammatische Raritätenkabinett*. Retrieved August 15, 2006 from < http://ling.uni-konstanz.de:591/universals/introrara.html >

Raben, J. 1987. Computers and the Humanities: some Historical Considerations. In Zampolli, A. (ed.). *Linguistica Computazionale, Volumi IV-V: Studies in Honor of Roberto Busa S.J*. Pisa: Giardini Editori. 225-230.

Raimondo, F. 2001. *Il lessico della parlata Arbëreshe di Acquaformosa*. Milano: Istituto Lombardo di Scienze e Lettere.

Ringe, D., T. Warnow, and A. Taylor. 2002. Indo-European and Computational Cladistics. In *Transactions of the Philological Society, 100(1)*. 59-129.

Riza, S. 2002. *Vepra 3*. Prishtinë: Akademia e Shkencave dhe e Arteve e Kosovës.

Sakita, T. 2002. *Reporting Discourse, Tense, and Cognition*. Amsterdam: Elsevier.

Scutari, P. 2002. *Il lessico della parlata arbëreshe di San Costantino Albanese*. Rende: Università della Calabria.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: University Press.

Smith, K. 2001. The role of frequency in the specialization of the English anterior. In Bybee and Hopper (2001:361-382).

Sturtevant, E. 1951. *A Comparative Grammar of the Hittite Language*. New Haven: Yale University Press.

Tanz, C. 1980. *Studies in the Acquisition of Deictic Terms*. Cambridge: Cambridge University Press.

Turano, G. 2001. *Tratti linguistici e culturali dell'Arbëria Crotonese*. Rende: Università della Calabria.

Trix, F. 1999. The Stamboul Alphabet of Shemseddin Sami Bey: Precursor to Turkish Script Reform. In *International Journal of Middle East Studies, Vol. 31, No. 2*. 255-272.

Xhuvani, A. and E. Çabej. 1976. Parashtesat e gjuhës shqipe. In Çabej, E. *Studime Gjuhësore III*. Prishtinë: Rilindja.

Žic Fuchs, M. 1996. 'Here' and 'there' in Croatian: a case study of an urban standard variety. In Pütz, M. and R. Driven (eds.). *The construal of space in language and thought*. Berlin – New York: Mouton de Gruyter. 49-62.

Appendix: Verb Actionality

		këtu				aty				atje							
Verb	English	#	f	F	ratio	#	f	F	ratio	#	f	F	ratio	dur	dyn	hom	
afroj	to approach					1	6	43	7.17					+	+	-	accomplishment
arrij	to arrive	1	40	5604	140.10	3	81	7811	96.43	1	19	804	42.32	-	+	-	achievement
banoj	to reside	1	12	276	23.00	3	24	816	34.00					+	-	+	stative
bëj	to make	3	154	13790	89.55									+	+	+	activity
dërgoj	to send	1	8	330	41.25					1	8	330	41.25	-	+	-	achievement
dua	to want	3	84	4068	48.43					1	12	1154	96.17	+	-	+	stative
flas	to speak	2	24	1091	45.46									+	+	+	activity
fle	to sleep					1	7	170	24.29					+	-	+	stative
fsheh	to hide					1	16	1226	76.63					+	-	+	stative
fus	to insert	1	19	2026	106.63									-	+	-	achievement
gjej	to find	1	6	306	51.00	5	105	7455	71.00	3	52	6114	117.58	-	+	-	activity
gjendem	to be found					4	37	1853	50.08					+	-	+	stative
harroj	to forget	2	28	1502	53.64	1	7	110	15.71					-	+	-	achievement
hyj	to enter					1	12	903	75.25					-	+	-	achievement
iki	to leave	2	24	1040	43.33									-	+	-	activity
jam	to be	7	438	20234	46.20	4	190	20283	106.75	5	172	26962	156.76	+	-	+	stative
jap	to give	1	10	594	59.40									-	+	-	achievement
jetoj	to live	4	51	3105	60.88	3	69	5157	74.74	4	66	3875	58.71	+	+	+	activity
kaloj	to pass					2	28	2016	72.00					-	+	-	achievement
kam	to have	4	430	37428	87.04	1	195	35770	183.44	4	139	22630	162.81	+	-	+	stative
kthehem	to return	1	6	222	37.00					1	9	222	24.67	-	+	-	achievement
kujtoj	to remember	2	44	641	14.57	2	28	1865	66.61					+	+	+	activity
kuptoj	to understand	1	13	292	22.46									+	-	+	stative
largohem	to leave									1	6	34	5.67	+	+	-	accomplishment
lë	to leave					3	87	8602	98.87					-	+	-	achievement
lind	to be born	1	25	1960	78.40	1	14	693	49.50					-	-	+	achievement
llogaris	to calculate	1	23	317	13.78									+	+	-	accomplishment
mbaroj	to complete	4	73	1962	26.88									-	+	-	achievement
mbërrij	to arrive	1	8	456	57.00	2	53	2793	52.70	3	55	2505	45.55	-	+	-	achievement
mbetem	to be left					1	44	4356	99.00	1	12	1049	87.42	+	-	+	stative
mbij	to sprout									1	6	11	1.83	-	+	-	achievement
mendoj	to think	1	36	3979	110.53	1	9	338	37.56					+	+	+	activity
merrem	to deal with	1	6	211	35.17									+	+	-	accomplishment
mungoj	to be absent	1	13	1037	79.77									+	-	+	stative
ndalem	to be stopped	3	24	341	14.21									-	+	-	achievement
ndërtoj	to build					1	6	88	14.67					+	+	-	accomplishment
ndihem	to feel	1	7	248	35.43									+	-	+	stative
ndodhem	to be found	1	12	56	4.67	4	185	7463	40.34	4	52	4622	88.88	+	-	+	stative
nis	to begin									1	7	274	39.14	-	+	-	achievement
njoftoj	to announce					1	6	193	32.17					-	+	-	achievement
pëlqej	to like	1	15	745	49.67									+	-	+	stative
përfshij	to include	4	423	4407	10.42									-	+	-	achievement
përfundoj	to conclude	1	19	975	51.32									-	+	-	achievement
përgjigjem	to answer	1	13	1036	79.69									+	+	-	accomplishment
përjashtoj	to exclude	3	37	1008	27.24									-	+	-	achievement
përmend	to mention	5	114	2324	20.39									-	+	-	achievement
pi	to drink					1	7	160	22.86					+	+	-	accomplishment
pres	to wait	2	14	570	40.71					1	21	746	35.52	+	-	+	stative
punoj	to work	2	28	2307	82.39	1	9	491	54.56	5	88	6139	69.76	+	+	+	activity
qëndroj	to stand	4	40	1956	48.90	6	82	4972	60.63	5	63	3837	60.90	+	+	+	activity
rri	to stay	4	51	1033	20.25	3	38	1268	33.37	2	20	956	47.80	+	+	+	activity
sëmurem	to get sick	1	6	78	13.00									-	+	-	achievement
shikoj	to look	1	7	296	42.29					1	7	296	42.29	+	+	+	activity
shkel	to step on									1	6	287	47.83	-	+	-	achievement
shkoj	to go	1	6	156	26.00	4	95	7593	79.93	15	373	14404	38.62	+	+	+	activity
shoh	to see	2	28	1161	41.46	1	6	139	23.17					+	-	+	stative

shtoj	to add	4	60	2594	43.23									-	+	-	achievement
shtroj	to lay down									1	13	953	73.31	+	+	-	accomplishment
sjell	to bring	2	25	1308	52.32									+	+	-	accomplishment
strehoj	to shelter					1	10	365	36.50					+	+	-	accomplishment
studioj	to study									1	6	49	8.17	+	+	+	activity
them	to say	4	90	7723	85.81					2	25	2041	81.64	+	+	+	activity
thërras	to call									1	7	128	18.29	+	+	+	activity
ul	to lower					1	16	1467	91.69					-	+	-	achievement
vdes	to die	1	9	636	70.67									+	+	-	accomplishment
vij	to come	14	436	18354	42.10	1	22	2195	99.77	1	6	91	15.17	+	+	+	activity
vlej	to be worth	1	15	775	51.67									+	-	+	stative
zhvillohem	to get developed					1	8	378	47.25					+	+	+	activity

		këtu				**aty**				**atje**			
	statives	25	1081	66890	61.88	20	704	72076	102.38	16	408	57163	140.11
	activities	45	997	55777	55.94	27	465	31712	68.20	40	713	37930	53.20
	achievements	34	900	26383	29.31	15	304	24588	80.88	9	110	4433	40.30
	accomplishments	5	63	2472	39.24	4	29	656	22.62	2	19	987	51.95